УДК 519.613
ББК 22.193
В 31

Федеральная целевая программа «Культура России»
(подпрограмма «Поддержка полиграфии и книгоиздания России»)

Рецензенты:

кафедра вычислительной математики Удмуртского государственного университета (зав. кафедрой — д-р физ.-мат. наук, проф. Г.Г. Исламов);
чл.-корр. РАН, проф. В.В. Васин

Вержбицкий, В.М.
В 31 Основы численных методов: Учебник для вузов/В.М. Вержбицкий. — М.: Высш. шк., 2002. — 840 с.: ил.

ISBN 5-06-004020-8

В книге систематически излагаются численные методы решения основных задач алгебры, математического анализа и дифференциальных уравнений (обыкновенных и с частными производными). Теоретический материал широко иллюстрируется таблицами, рисунками, примерами и библиографическими ссылками. В каждой главе даются упражнения для самостоятельной работы. Одно из двух приложений содержит образцы постановок лабораторных работ по всему курсу численных методов, в другом приводятся элементарные сведения из функционального анализа.

Для студентов математических и инженерных специальностей вузов. Может быть полезна широкому кругу читателей, интересующихся вычислительной математикой.

УДК 519.613
ББК 22.193

ISBN 5-06-004020-8

Оглавление

50-летию Ижевского государственного технического университета посвящается

ПРЕДИСЛОВИЕ

Предлагаемая книга — плод более чем тридцатилетнего опыта преподавательской работы автора на кафедре прикладной математики и информатики Ижевского государственного технического университета (бывшего механического института). В нее включен материал двух предыдущих книг автора «Численные методы (линейная алгебра и нелинейные уравнения)» и «Численные методы (математический анализ и обыкновенные дифференциальные уравнения)», выпущенных издательством «Высшая школа» соответственно в 2000 и 2001 гг., который дополнен кратким изложением методов решения уравнений с частными производными. Если вторая из названных книг помещена сюда практически без изменений, то первая значительно переработана. Кроме исправления замеченных опечаток и пополнения множества примеров и упражнений, облегчающих понимание изучаемых методов, новым в этой части является следующее.

• Первая глава дополнена параграфом 1.7, где можно получить первые представления о корректных и некорректных постановках задач и методах регуляризации (далее, в гл. 13, на этой основе показывается регуляризуемость задачи численного дифференцирования).

• В гл. 5 (§ 5.7) излагаются совершенно новые модификации методов Ньютона и секущих решения нелинейных скалярных уравнений, так называемые полюсные методы, имеющие зачастую более высокую скорость сходимости, чем их классические прообразы. Изложение этих методов опирается на совместные с автором исследования М.Ю.Петрова и Н.А.Рычиной.

• В § 7.5 главы 7 (соответствующей гл. 6 первой книги) рассматривается обобщение полюсного метода Ньютона на случай систем нелинейных конечномерных уравнений, предложенное автором и М.Ю.Петровым. Кроме того, в этой главе (§ 7.4) описан метод секущих Бройдена, являющийся одним из наибо-

лее эффективных (по требуемому числу вычислений функций) методов решения нелинейных систем и, в то же время, практически неупоминаемый в отечественной литературе.

Цель, которую ставил перед собой автор при написании данной книги — это по возможности полно охватить традиционный учебный курс вычислительной математики, причем так, чтобы книга представляла интерес для как можно более широкой читательской аудитории с разными уровнями математической подготовки, со своими представлениями о необходимой строгости и достаточной полноте изложения отдельных фактов, с разным количеством часов, отводимых учебными планами под изучение основ численных методов и, вообще, с разной потребностью в предмете изучения. Думается, что достижению этой цели должны способствовать следующие обстоятельства.

• Автор старался донести до читателя прежде всего идею каждого метода решения той или иной задачи. При этом некоторые методы полностью обосновывались и доводились до более-менее подробного алгоритма, какие-то методы после рассмотрения идеи сразу снабжались алгоритмами, а знакомство с какими-то из методов оставалось на идейном уровне. Если метод не рассматривался подробно, то обязательно делались ссылки на литературные источники (как правило, это учебные пособия или монографии), где можно найти упоминающийся материал. Наиболее полно здесь освещены численные методы решения конечномерных уравнений и задач математического анализа и менее полно (в смысле обоснования) методы решения дифференциальных уравнений, особенно с частными производными.

• Считая, что изучение основ численных методов почти немыслимо без вычислений, в одном из приложений автор приводит возможные постановки заданий для лабораторных работ по курсу, что вместе с сопровождающими изложение методов примерами, а также с помещенными в конце каждой главы упражнениями должно способствовать осмысленному усвоению изучаемого материала. В списке литературы можно найти сведения о задачниках по численным методам (см. [15, 65, 87, 89, 142, 162, 163]), которые полезно иметь в виду при организации практических и лабораторных занятий.

• Хотя автор старался не погружаться в функциональный анализ, в определенных местах обойтись совсем без его некоторых первичных понятий не смог (или не захотел, ибо взглянув на некоторые вещи чуть-чуть сверху, мы видим их гораздо отчетливее). Чтобы такие места без особого труда воспринимались читателями, совсем незнакомыми с функциональным анализом, в приложении очень кратко даны необходимые сведения из этой математической дисциплины.

В какой мере автору удалось достичь поставленной цели, судить читателю. Автор понимает, что в своей попытке «объять необъятное» он создал труд, который одним покажется излишне подробным, но недостаточно аргументированным, другим — наоборот, перегруженным выводами формул. Хотелось бы надеяться, что многим из тех, кто будет держать в руках эту книгу, по крайней мере, в какой-то момент она окажется нужной и полезной.

Автор выражает глубокую благодарность Ученому Совету и ректору ИжГТУ проф. И.В.Абрамову за предоставление времени для написания книги, коллегам по кафедре ПМИ и ее заведующему доц. А.А.Айзиковичу за содействие в подготовке рукописи, проф. А.Л.Тептину за внимательное ее прочтение и ценные замечания, А.В.Чуракову и М.Ю.Петрову за компьютерный набор текста, формул и рисунков, а также Ю.В.Гаврину за приведение набора к единой форме и макетирование книги. Автор искренне признателен чл.-корр. РАН В.В.Васину и проф. Г.Г.Исламову за проделанную работу по рецензированию книги. Особо благодарен автор сыну П.В.Вержбицкому, оказавшему всестороннюю помощь на разных этапах работы над книгой.

Несмотря на то, что к процессу подготовки книги к изданию причастно много людей, автор не желает переложить на других и доли ответственности за ошибки и опечатки, без которых, к сожалению, не обходится практически ни одно издание, и будет благодарен всем, кто сообщит о таковых автору на E-mail pmi@istu.udm.ru или vervm@verba.udm.ru.

Автор

ГЛАВА 1 ОБ УЧЕТЕ ПОГРЕШНОСТЕЙ ПРИБЛИЖЕННЫХ ВЫЧИСЛЕНИЙ

Рассматривается круг вопросов, связанных с учетом погрешностей, появление которых неизбежно при численном анализе математических моделей, в частности, при решении линейных алгебраических систем и нелинейных скалярных уравнений. Обращается внимание на заведомо приближенный характер компьютерных операций над действительными числами. Приводятся примеры задач и методов, чрезмерно чувствительных к ошибкам исходных данных и к погрешностям арифметических действий. Даются первые представления о корректных и некорректных (по Адамару и по Тихонову) задачах, регуляризирующем алгоритме и о методе α-регуляризации Тихонова.

1.1. ОБЩАЯ ФОРМУЛА ДЛЯ ОЦЕНКИ ГЛАВНОЙ ЧАСТИ ПОГРЕШНОСТИ

При численном решении математических и прикладных задач почти неизбежно появление на том или ином этапе их решения погрешностей следующих трех типов.

1) ***Погрешность задачи.*** Она связана с приближенным характером исходной содержательной модели (в частности, с невозможностью учесть все факторы в процессе изучения моделируемого явления), а также ее математического описания, параметрами которого служат обычно приближенные числа (например, из-за принципиальной невозможности выполнения абсолютно точных измерений). Для вычислителя погрешность задачи следует считать ***неустранимой*** (безусловной), хотя постановщик задачи иногда может ее изменить.

2) ***Погрешность метода.*** Это погрешность, связанная со способом решения поставленной математической задачи и появляющаяся в результате подмены исходной математической модели другой или конечной последовательностью других, например, линейных моделей. При создании численных методов закладывается возможность отслеживания таких погрешностей и доведения их до сколь угодно малого уровня. Отсюда естественно отношение к погрешности метода как к ***устранимой*** (или условной).

3) ***Погрешность округлений*** (погрешность действий). Этот тип погрешностей обусловлен необходимостью выполнять арифметические операции над числами, усеченными до количества разрядов, зависящего от применяемой вычислительной техники (если, разумеется, не используются специальные программные средства, реализующие, например, арифметику рациональных чисел).

Все три описанных типа погрешностей в сумме дают ***полную погрешность*** результата решения задачи. Поскольку первый тип погрешностей не находится в пределах компетенции вычислителя, для него он служит лишь ориентиром точности, с которой следует рассчитывать математическую модель. Нет смысла решать задачу существенно точнее, чем это диктуется неопределенностью исходных данных. Таким образом, погрешность метода подчиняют погрешности задачи. Наконец, при выводе оценок погрешностей численных методов обычно исходят из предположения, что все операции над числами выполняются точно. Это означает, что погрешность округлений не должна существенно отражаться на результатах реализации методов, т.е. должна подчиняться погрешности метода. Влияние погрешностей округлений не следует упускать из вида ни на стадии отбора и алгоритмизации численных методов, ни при выборе вычислительных и программных средств, ни при выполнении отдельных действий и вычислении значений функций.

Рассмотрим некоторые возможные подходы к учету погрешностей действий ([3, 20, 25, 44, 61, 71, 73, 99] и др.).

Пусть A и a — два «близких» числа; условимся считать A — точным, a — приближенным.

Величина $\Delta a := |A - a|$ называется ***абсолютной погрешностью*** приближенного числа a, а $\delta a := \dfrac{\Delta a}{|a|}$ — его ***относительной погрешностью****). Числа Δ_a и δ_a такие, что $\Delta_a \ge \Delta a$ и $\delta_a = \dfrac{\Delta_a}{|a|} \ge \delta a$, называются ***оценками*** или ***границами*** абсолютной и относительной погрешностей соответственно (к Δ_a и δ_a часто применяют также термин ***«предельные погрешности»***). Так как обычно истинные погрешности не известны, то там, где

---

*) 1. Символ «:=» нами используется в двух близких смыслах:

а) положить по определению, задать, выбрать;

б) присвоить, т.е. записать в ячейку памяти компьютера, определяемую идентификатором в левой части, значение, задаваемое правой частью.

2. Возможно, чаще относительной погрешностью называют $\dfrac{\Delta a}{|A|}$.

Использование символа «Δ» в дальнейшем может быть двояким: как для обозначения абсолютной погрешности, так и для обозначения приращения переменной, что будет ясно из контекста либо специально оговорено.

не может возникнуть недоразумений, будем иногда называть Δ_a и δ_a просто абсолютной и относительной погрешностями.

Поставим вопрос о грубом оценивании погрешности результата вычисления значения дифференцируемой функции $u = f(x_1, x_2, \ldots, x_n)$ приближенных аргументов $x_1, x_2, \ldots, x_n$, если известны границы их абсолютных погрешностей $\Delta_{x_1}, \Delta_{x_2}, \ldots, \Delta_{x_n}$, соответственно. В этом случае точные значения аргументов $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ лежат соответственно на отрезках $[x_1 - \Delta_{x_1}, x_1 + \Delta_{x_1}]$, $[x_2 - \Delta_{x_2}, x_2 + \Delta_{x_2}]$, ..., $[x_n - \Delta_{x_n}, x_n + \Delta_{x_n}]$, а точная абсолютная погрешность результата $u = f(x_1, x_2, \ldots, x_n)$ есть

$$\Delta u = |f(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n) - f(x_1, x_2, \ldots, x_n)|$$

— модуль полного приращения функции. Главной, т.е. линейной частью этого приращения является, как известно, полный дифференциал du. Таким образом, имеем:

$$\Delta u \approx |du| = \left|\sum_{i=1}^{n} \frac{\partial u}{\partial x_i} dx_i\right| \le \sum_{i=1}^{n} \left|\frac{\partial u}{\partial x_i}\right| \cdot |\tilde{x}_i - x_i| \le \sum_{i=1}^{n} \left|\frac{\partial u}{\partial x_i}\right| \cdot \Delta_{x_i},$$

т.е. за границу абсолютной погрешности результата приближенно может быть принята величина

$$\Delta_u = \sum_{i=1}^{n} \left|\frac{\partial u}{\partial x_i}\right| \cdot \Delta_{x_i}. \qquad (1.1)$$

Отсюда легко получается формула приближенной оценки относительной погрешности значения u:

$$\delta_u = \frac{\Delta_u}{|u|} = \sum_{i=1}^{n} \left|\frac{\partial u}{\partial x_i}\right| \cdot \frac{\Delta_{x_i}}{|u|} = \sum_{i=1}^{n} \left|\frac{\partial u}{u \partial x_i}\right| \cdot \Delta_{x_i} = \sum_{i=1}^{n} \left|\frac{\partial \ln u}{\partial x_i}\right| \cdot \Delta_{x_i}. \qquad (1.2)$$

Как частные случаи формул (1.1), (1.2) (точных для функций, линейных относительно x_i или $\ln x_i$, соответственно) можно получить известные правила оценивания погрешностей результатов арифметических действий.

Действительно, пусть $u = \pm x_1 \pm x_2 \pm \ldots \pm x_n$. Тогда $\left|\frac{\partial u}{\partial x_i}\right| = 1$ и $\Delta_{\Sigma(\pm x_i)} = \sum_{i=1}^{n} 1 \cdot \Delta_{x_i} = \sum_{i=1}^{n} \Delta_{x_i}$, т.е. при сложении и вычитании приближенных чисел их предельные абсолютные погрешности складываются.

Пусть теперь $u = x_1 \cdot x_2 \cdot \ldots \cdot x_n$, где можно считать все сомножители положительными. Так как $\ln u = \ln x_1 + \ln x_2 + \ldots + \ln x_n$ и $\frac{\partial \ln u}{\partial x_i} = \frac{1}{x_i}$, то, согласно (1.2),

$$\delta_{\Pi x_i} = \sum_{i=1}^{n} \frac{1}{x_i} \Delta_{x_i} = \sum_{i=1}^{n} \delta_{x_i}. \qquad (1.3)$$

Если же $u = \frac{x_1}{x_2}$, где $x_1, x_2 > 0$, то $\ln u = \ln x_1 - \ln x_2$, $\frac{\partial \ln u}{\partial x_i} = \frac{1}{x_i}$ и, значит,

$$\delta_{x_1/x_2} = \frac{\Delta_{x_1}}{x_1} + \frac{\Delta_{x_2}}{x_2} = \delta_{x_1} + \delta_{x_2}.$$

Последнее вместе с (1.3) означает известный результат о сложении предельных относительных погрешностей при умножении и делении приближенных чисел.

Возвращаясь к сложению, рассмотрим относительную погрешность суммы n положительных приближенных чисел $x_1, x_2, \ldots, x_n$, имеющих границы относительных погрешностей $\delta_{x_1}, \delta_{x_2}, \ldots, \delta_{x_n}$ соответственно:

$$\delta(x_1 + x_2 + \ldots + x_n) = \frac{\Delta(x_1 + x_2 + \ldots + x_n)}{x_1 + x_2 + \ldots + x_n} \le$$

$$\le \frac{\Delta_{x_1 + x_2 + \ldots + x_n}}{x_1 + x_2 + \ldots + x_n} = \frac{\Delta_{x_1} + \Delta_{x_2} + \ldots + \Delta_{x_n}}{x_1 + x_2 + \ldots + x_n} =$$

$$= \frac{x_1 \delta_{x_1} + x_2 \delta_{x_2} + \ldots + x_n \delta_{x_n}}{x_1 + x_2 + \ldots + x_n} \le \frac{x_1 \delta^* + x_2 \delta^* + \ldots + x_n \delta^*}{x_1 + x_2 + \ldots + x_n} = \delta^*,$$

где $\delta^* := \max_i \delta_{x_i}$. Полученное неравенство говорит о том, что относительная погрешность суммы n положительных приближенных чисел не превосходит максимальной относительной погрешности слагаемых.

С вычитанием приближенных чисел дело обстоит хуже: оценка

$$\delta_{x_1 - x_2} = \frac{\Delta_{x_1 - x_2}}{|x_1 - x_2|} = \frac{\Delta_{x_1} + \Delta_{x_2}}{|x_1 - x_2|}$$

относительной погрешности разности $x_1 - x_2$ двух приближен-

ных положительных чисел указывает на возможность сильного возрастания погрешности при $x_1 - x_2 \to 0$. В этом случае говорят о потере точности при вычитании близких чисел.

Часто возникает ***обратная задача теории погрешностей***: какой точности данные нужно подать на вход, чтобы на выходе получить результат заданной точности? Применительно к поставленной выше прямой задаче оценивания погрешности результата вычисления значения функции при заданных оценках погрешностей аргументов обратная задача заключается в оценивании величин Δx_i (или δx_i) по известной величине Δu. Для случая дифференцируемой функции одной переменной грубое решение обратной задачи тривиально: если $y = f(x)$, то $\Delta y \approx |dy| = |f'(x)|\Delta x$, откуда $\Delta x \approx \dfrac{\Delta y}{|f'(x)|}$. Для функции большего числа переменных обратная задача, вообще говоря, некорректна. Нужны дополнительные условия. Например, применяют ***принцип равных влияний***, состоящий в предположении, что частные дифференциалы $\left|\dfrac{\partial u}{\partial x_i}\right|\Delta_{x_i}$ в (1.1) одинаково влияют на погрешность значения функции; тогда

$$\Delta_u = n\left|\frac{\partial u}{\partial x_i}\right|\Delta_{x_i}, \quad \text{откуда} \quad \Delta_{x_i} = \frac{\Delta_u}{n\left|\dfrac{\partial u}{\partial x_i}\right|}.$$

В качестве другого довольно естественного допущения можно принять равенство относительных погрешностей всех аргументов, т.е. считать $\delta_{x_i} = \dfrac{\Delta_{x_i}}{|x_i|} = p$ при всех $i = 1, 2, \ldots, n$. Тогда $\Delta_{x_i} = p|x_i|$ и, значит, $\Delta_u = p\sum_{i=1}^{n}\left|\dfrac{\partial u}{\partial x_i}x_i\right|$. Из последнего равенства получаем величину $p = \dfrac{\Delta_u}{\sum_{i=1}^{n}\left|x_i\dfrac{\partial u}{\partial x_i}\right|}$ (характеризующую относительный уровень точности задания аргументов), на основе которой за границы абсолютных погрешностей аргументов принимаем $\Delta_{x_i} = \dfrac{|x_i|\Delta_u}{\sum_{i=1}^{n}\left|x_i\dfrac{\partial u}{\partial x_i}\right|}$. Имеются и другие, более сложные подходы к решению обратной задачи (см., например, [13]).

1.2. СТАТИСТИЧЕСКИЙ И ТЕХНИЧЕСКИЙ ПОДХОДЫ К УЧЕТУ ПОГРЕШНОСТЕЙ ДЕЙСТВИЙ

Рассмотренный выше ***аналитический*** (или ***классический***) ***способ учета погрешностей действий***, предполагающий точное оценивание погрешностей, основанное либо на приведенных в предыдущем параграфе правилах подсчета погрешностей арифметических действий, либо на параллельной работе с верхними и нижними границами исходных данных, имеет два существенных недостатка. Во-первых, этот способ чрезвычайно громоздок и не может быть рекомендован при массовых вычислениях. Во-вторых, он учитывает крайние, наихудшие случаи взаимодействия погрешностей, которые допустимы, но маловероятны. Ясно, что, например, при суммировании нескольких приближенных чисел (полученных в результате измерений, округлений или каким-либо другим путем) среди них почти наверное будут слагаемые как с избытком, так и с недостатком, т.е. произойдет частичная компенсация погрешностей. При больших количествах однотипных вычислений вступают в силу уже ***вероятностные*** или ***статистические законы*** формирования погрешностей результатов действий. Например, методами теории вероятностей показывается, что математическое ожидание абсолютной погрешности суммы n слагаемых с одинаковым уровнем абсолютных погрешностей, при достаточно большом n, пропорционально $\sqrt{n}$ ([20, 25, 61]). В частности, если $n > 10$ и все слагаемые округлены до m-го десятичного разряда, то для подсчета абсолютной погрешности суммы S применяют ***правило Чеботарева***

$$\Delta S \approx \sqrt{3n}\cdot 0.5\cdot 10^{-m}. \tag{1.4}$$

Различие в результатах классического и статистического подходов к оцениванию погрешности суммы рассмотрим на примере оценки погрешности среднего арифметического нескольких приближенных чисел.

Пусть $x = \frac{1}{n}(x_1 + \ldots + x_n)$ — среднее арифметическое n ($n > 10$) приближенных чисел (например, результатов измерений), имеющих одинаковый уровень абсолютных погрешностей $\Delta_{x_i} = 0.5 \cdot 10^{-m}$. Тогда классическая оценка абсолютной погрешности величины x есть

$$\Delta_x = \frac{1}{n}(\Delta_{x_1} + \ldots + \Delta_{x_n}) = \frac{1}{n} \cdot n \cdot 0.5 \cdot 10^{-m} = 0.5 \cdot 10^{-m} = \Delta_{x_i},$$

т.е. такая же, как и у исходных данных. В то же время по формуле (1.4) имеем

$$\Delta_x \approx \frac{1}{n}\sqrt{3n} \cdot 0.5 \cdot 10^{-m} = \sqrt{\frac{3}{n}} \cdot 0.5 \cdot 10^{-m} = \sqrt{\frac{3}{n}} \cdot \Delta_{x_i} \xrightarrow[n \to \infty]{} 0.$$

Как видим, применение правила Чеботарева приводит к естественному выводу о том, что арифметическое усреднение результатов измерений или наблюдений увеличивает точность, чего нельзя сказать на основе классической теории погрешностей.

Прямое применение вероятностно-статистических оценок погрешностей также является достаточно сложным делом и вряд ли может быть рекомендовано при рядовых массовых вычислениях. Однако именно такие оценки подкрепляют практические правила работы с приближенными числами, составляющие основу так называемого ***технического подхода***. Этот подход связывают с именем известного русского кораблестроителя, математика и механика академика А. Н. Крылова*). Согласно ***принципу А. Н. Крылова***, *приближенное число должно записываться так, чтобы в нем все значащие цифры, кроме последней, были верными и лишь последняя была бы сомнительна, и притом в среднем***) *не более чем на одну единицу*. Напомним, что ***значащими цифрами*** числа в его позиционной записи называются все его цифры, начиная с первой ненулевой слева. Значащую цифру приближенного числа называют ***верной***, если абсолютная погрешность числа не превосходит единицы разряда, в котором стоит эта цифра (или половины единицы; в этом случае иногда применяется термин *верная в узком смысле*).

Чтобы результаты арифметических действий, совершаемых над приближенными числами, записанными в соответствии с

---

*) Крылов Алексей Николаевич (1863–1945).

**) «В среднем» здесь понимается в вероятностном смысле.

принципом А. Н. Крылова, также соответствовали этому принципу, нужно придерживаться следующих простых правил [25, 44, 61, 98]:

1) *при сложении и вычитании приближенных чисел в результате следует сохранять столько десятичных знаков, сколько их в приближенном данном с наименьшим количеством десятичных знаков*;

2) *при умножении и делении в результате следует сохранять столько значащих цифр, сколько их имеет приближенное данное с наименьшим числом значащих цифр*;

3) *результаты промежуточных вычислений должны иметь один-два запасных знака (которые затем должны быть отброшены)*.

Таким образом, при техническом подходе к учету погрешностей приближенных вычислений предполагается, что в самой записи приближенного числа содержится информация о его точности. И хотя прямая выгода от применения приведенных правил работы с приближенными числами может быть получена лишь при ручном счете (не нужно оперировать с цифрами, не влияющими на информативную часть приближенного результата), их знание и понимание помогает правильной интерпретации компьютерных расчетов, а иногда и самой организации таковых.

1.3. ПОНЯТИЕ О ПОГРЕШНОСТЯХ МАШИННОЙ АРИФМЕТИКИ

Для представления вещественных чисел в компьютере применяют, в основном, два способа: с фиксированной и с плавающей запятой (точкой).

Пусть в основу запоминающего устройства машины положены однотипные физические устройства (базисные элементы), имеющие r устойчивых состояний (как правило, $r = 2, 8, 16$ и т.п.), причем каждому числу ставится в соответствие одинаковое количество k этих элементов и, кроме того, с помощью таких или более простых элементов может фиксироваться знак. Упорядоченные элементы образуют разрядную сетку машинного слова: в каждом разряде может быть записано одно из базисных чисел $0, 1, \ldots, r-1$ (одна из r «цифр» r-ичной системы счисления) и в специальном разряде отображен знак $+$ или $-$.

При записи числа с ***фиксированной запятой*** кроме упомянутых r параметров (основания системы счисления) и k (количества разрядов, отводимых под запись цифр числа) указывается

еще количество l разрядов, выделяемых под дробную часть числа. Таким образом, положительное вещественное число a, представляющее собой в r-ичной системе бесконечную, вообще говоря, непериодическую дробь, здесь будет отображено конечной последовательностью

$$\alpha_1\alpha_2\ldots\alpha_{k-l}\alpha_{k-l+1}\ldots\alpha_{k-1}\alpha_k,$$

где $\alpha_i \in \{0, 1, \ldots, r-1\}$, т.е. реализуется приближенное равенство

$$a \approx \operatorname{fix}(a) := \alpha_1 r^{k-l-1} + \alpha_2 r^{k-l-2} + \cdots + \alpha_{k-l} r^0 + \\ + \alpha_{k-l+1} r^{-1} + \cdots + \alpha_{k-1} r^{-(l-1)} + \alpha_k r^{-l}.$$

Диапазон представляемых таким способом чисел определяется числами с наибольшими цифрами во всех разрядах, т.е. наименьшим $-(r-1)(r-1)\cdots(r-1)$ и наибольшим $(r-1)(r-1)\cdots(r-1)$ числами, а ***абсолютная точность представления*** есть оценка величины $|a-\operatorname{fix}(a)|$, зависящая от способа округления: это r^{-l} при простом отбрасывании «хвоста» $\alpha_{k+1} r^{-(l+1)} + \alpha_{k+2} r^{-(l+2)} + \ldots$ числа a и половина этой величины при ***правильном округлении*** (т.е. при увеличении α_k на единицу, если $\alpha_{k+1} > \frac{r}{2}$). Заметим, что абсолютная точность представления вещественных чисел с фиксированной запятой одинакова в любой части диапазона. В то же время ***относительная точность***, т.е. оценка величины $\left|\frac{a-\operatorname{fix}(a)}{a}\right|$ (или $\left|\frac{a-\operatorname{fix}(a)}{\operatorname{fix}(a)}\right|$), очевидно, может значительно различаться в зависимости от того, берется a близким к нулю или к границе диапазона. Иными словами, вещественные числа с фиксированной запятой имеют равномерную абсолютную плотность распределения на всем отрезке вещественной оси, определяемом границами диапазона, и неравномерную, возрастающую к границам отрезка, относительную плотность распределения.

В основе значительно чаще употребляемого представления с ***плавающей запятой*** лежит следующая экспоненциальная форма записи вещественного числа:

$$a = M \cdot r^p,$$

где r — ***основание***, p — ***порядок***, а M такое, что $r^{-1} \le |M| < 1 \left(= r^0\right)$ — ***мантисса***. Если под мантиссу выделяется l

r-ичных элементов, а под порядок m, то в системе записи с плавающей запятой вещественное число a представляется конечным числом $\operatorname{fl}(a)$ (от англ. floating — «плавающий») вида

$$a \approx \operatorname{fl}(a) := \pm\left(\beta_1 r^{-1} + \beta_2 r^{-2} + \ldots + \beta_l r^{-l}\right) \cdot r^{\gamma},$$

где γ — целое число из промежутка $\left[-r^m,\ r^m - 1\right]$; $\beta_1 \in \{1, \ldots, r-1\}$; $\beta_i \in \{0, 1, \ldots, r-1\}$ $(i = 2, \ldots, l)$, т.е. ***машинное слово*** условно имеет структуру:

Знак порядка	Порядок	Знак мантиссы	Мантисса
	m разрядов		l разрядов

Числа $\pm r^{r^m}$ определяют границы допустимого числового диапазона. Более информативно здесь нужно говорить о диапазоне представимости положительных вещественных чисел, составляющем промежуток $\left[r^{-r^m},\ r^{r^m-1}\right]$. Левую и правую границы этого отрезка называют соответственно ***машинным нулем*** и ***машинной бесконечностью***, так как числа из промежутка $\left[-r^{-r^m}, r^{-r^m}\right]$ машина заменяет нулем, а числа, лежащие за пределами промежутка $\left[-r^{r^m-1},\ r^{r^m-1}\right]$, она не воспринимает (без специальных ухищрений).

Важной характеристикой является число ε, называемое ***машинный эпсилон*** и обозначаемое обычно идентификатором ***macheps***. Эта характеристика определяется как расстояние между единицей и ближайшим следующим за ней числом системы машинных чисел с плавающей запятой. Так как

$$1 = \left(1 \cdot r^{-1} + 0 \cdot r^{-2} + \ldots + 0 \cdot r^{-l} + \ldots\right) \cdot r^1,$$

а следующее за 1 машинное число есть

$$\left(1 \cdot r^{-1} + 0 \cdot r^{-2} + \ldots + 0 \cdot r^{-(l-1)} + 1 \cdot r^{-l}\right) \cdot r^1 = \operatorname{fl}(1+\varepsilon),$$

то за *macheps* можно принять величину

$$\varepsilon = 1 \cdot r^{-l} \cdot r^1 = r^{1-l}.$$

Это число непосредственно связано с относительной погрешностью представления чисел в системе с плавающей запятой. Имеем:

$$\left|\frac{a - \operatorname{fl}(a)}{a}\right| = \frac{\beta_{\ell+1} r^{-(l+1)} + \beta_{\ell+2} r^{-(l+2)} + \cdots}{\beta_1 r^{-1} + \beta_2 r^{-2} + \cdots} \le \frac{1 \cdot r^{-l}}{\beta_1 \cdot r^{-1}} \le r^{1-l} = \varepsilon. \quad (1.5)$$

Таким образом, *машинный эпсилон служит мерой относительной точности представления вещественных чисел*, причем эта точность одинакова в любой части числового диапазона и зависит лишь от числа r-ичных разрядов, отводимых под мантиссу числа. В то же время оценка абсолютной погрешности

$$|a - \mathrm{fl}(a)| \le |a| \cdot r^{1-l}$$

показывает, что расстояние между вещественными числами и конечными приближениями к ним в системе с плавающей запятой неодинаковы в разных частях числового диапазона: *абсолютная плотность машинных чисел больше вблизи нуля при одинаковой относительной плотности их распределения*.

Замечание 1.1. Если трактовать *macheps* как минимальное положительное действительное число, прибавление которого к 1 дает следующее за 1 число с плавающей запятой, то, очевидно, при правильном округлении значение *macheps* будет в два раза меньшим. Действительно, полагая $\varepsilon = \frac{1}{2} r^{1-l} = \frac{r}{2} \cdot r^{-(l+1)} \cdot r^1$, получаем

$$1 + \varepsilon = \left(1 \cdot r^{-1} + 0 \cdot r^{-2} + \ldots + 0 \cdot r^{-l} + \frac{r}{2} \cdot r^{-(l+1)}\right) \cdot r^1$$

и, значит, $\mathrm{fl}(1+\varepsilon) = 1 + r^{1-l} > 1$. «Мера дискретности» множества машинных чисел, как видим, остается той же: r^{1-l}.

Замечание 1.2. Величина *macheps* служит оценкой относительной точности представления вещественного числа a при условии, что $|a| > r^{-r^m}$. Если же $a \in \left[-r^{-r^m},\ r^{-r^m}\right]$, то $\mathrm{fl}(a) \equiv 0$ и, значит, относительная погрешность

$$\left|\frac{a - \mathrm{fl}(a)}{a}\right| \equiv 1,$$

т.е. является постоянной достаточно большой величиной, в то время как абсолютная погрешность не превосходит величины r^{-r^m}.

Приведем значения введенных выше теоретических параметров для нескольких типов отечественных ЭВМ (электронно-вычислительных машин) 60-80 гг.*) (реальные параметры, естественно, могут незначительно отличаться от приводимых) [3, 6, 50, 54, 158].

---

*) В качестве упражнения читателю предлагается составить несложную программу, которая позволит получить соответствующие параметры используемого им современного компьютера. В готовом виде такую программу можно обнаружить в книге [68].

Так, для записи числа в 48-разрядном машинном слове БЭСМ-6 40 двоичных разрядов выделяются под мантиссу, 6 — под порядок и 2 — под знаки мантиссы (т.е. числа) и порядка. Отсюда, принимая $r = 2$, $l = 40$, $m = 6$, получаем, что точность представления чисел с плавающей запятой на БЭСМ-6 не хуже $2^{-39} \left(\approx 10^{-12}\right)$, граница машинного нуля $2^{-64} \left(\approx 10^{-19}\right)$, машинной бесконечности $2^{63} \left(\approx 10^{19}\right)$.

Машинное слово СМ ЭВМ имеет 32 двоичных разряда, из которых под мантиссу выделяется 24, а под порядок — 7. Зная параметры $r = 2$, $l = 24$, $m = 7$, получаем $macheps = 2^{-23} \approx 10^{-7}$, машинные нуль и бесконечность $\approx 10^{\mp 38}$.

На ЕС ЭВМ используется представление вещественных чисел по основанию $r = 16$. Эти машины имеют относительную точность представления $\approx 10^{-7}$ и диапазон для положительных чисел $\approx 10^{-77} \div 10^{76}$.

Практически любая машина ЕС ЭВМ рассчитана на то, чтобы выделить под запись числа двойное машинное слово, что позволяет более чем вдвое увеличивать точность представления. Двойная точность предусматривается также многими языками программирования.

Обращаясь к арифметическим операциям над машинными числами, прежде всего заметим, что они утрачивают привычные свойства. Особенно это касается свойств ассоциативности и дистрибутивности, нарушаемых при выполнении арифметических операций на любых ЭВМ (сохранение или несохранение свойства коммутативности связывают со способом округления чисел). Так, весьма утрированный пример сравнения выражения

$$\left(r^{D/2} \cdot r^{3D/4}\right) \cdot r^{-D/2} \quad \text{с выражением} \quad r^{D/2} \cdot \left(r^{3D/4} \cdot r^{-D/2}\right),$$

где D таково, что r^D — правая граница числового диапазона, показывает существенность способа расстановок скобок при умножении: в первом случае машина выдаст сообщение о переполнении, в силу

$$r^{D/2} \cdot r^{3D/4} > r^D,$$

а во втором — будет получен правильный результат. Легко также представить ситуацию с тремя положительными числами a, b, c, когда расставляя по-разному скобки в выражении $a + b - c$, будем получать (или не получать вовсе) разные результаты.

Изучение погрешностей результатов арифметических операций над числами с плавающей запятой производят с помощью

представления

$$\mathrm{fl}(a)=a(1+\delta), \quad \text{где} \quad |\delta|\le macheps \tag{1.6}$$

(чтобы убедиться в справедливости (1.6), достаточно ввести δ равенством $\delta=\frac{\mathrm{fl}(a)-a}{a}$, равносильным фигурирующему в (1.6), и воспользоваться доказанным в (1.5) неравенством $|\delta|\le macheps$). Принимая во внимание, что операции над двумя машинными числами a и b осуществляются точно (здесь используется двойная длина машинного слова), после чего производится округление, результат любой арифметической операции $\otimes$ также может быть записан в виде

$$\mathrm{fl}(a\otimes b)=(a\otimes b)(1+\delta_1), \tag{1.7}$$

где $|\delta_1|\le\varepsilon$ (за исключением особых случаев).

Пусть складываются последовательно три положительных числа a_1, a_2, a_3. Тогда, согласно (1.7),

$$\mathrm{fl}(a_1+a_2)=(a_1+a_2)(1+\delta_1),$$

где $|\delta_1|\le\varepsilon$;

$$\mathrm{fl}((a_1+a_2)+a_3)=((a_1+a_2)(1+\delta_1)+a_3)(1+\delta_2)=$$
$$=(a_1+a_2)(1+\delta_1)(1+\delta_2)+a_3(1+\delta_2),$$

где $|\delta_i|\le\varepsilon$ $(i=1,\ 2)$.

Заменяя здесь δ_i бóльшим значением ε, получим оценку абсолютной погрешности суммы трех слагаемых:

$$|\,\mathrm{fl}(a_1+a_2+a_3)-(a_1+a_2+a_3)|\le 2(a_1+a_2)\varepsilon+a_3\varepsilon+(a_1+a_2)\varepsilon^2.$$

Обращает на себя внимание неравноправность слагаемых в образовании погрешности суммы: меньшую роль в ней играет последнее слагаемое. Природа этого факта очевидна: первые слагаемые неявно (в просуммированном виде) участвуют в процессе каждого последующего сложения.

Если пренебречь степенями ε выше первой, то для суммы n положительных чисел a_i нетрудно получить [6] приближенную оценку абсолютной погрешности вида*)

$$\left|\mathrm{fl}\left(\sum_{i=1}^{n}a_i\right)-\sum_{i=1}^{n}a_i\right|\lesssim|(n-1)(a_1+a_2)+(n-2)a_3+\ldots+2a_{n-1}+a_n|\cdot\varepsilon$$

*) Здесь и далее знак «$\lesssim$» используется для обозначения неравенства в смысле главных (линейных) частей.

при последовательном суммировании, начинающемся с a_1. Очевидно, чтобы эта погрешность была минимальной, *последовательность чисел нужно суммировать в порядке возрастания членов*. Только за счет этого можно добиться уменьшения погрешности, как показано в [42], в $n/\log_2 n$ раз.

На основе изучения погрешности произведения нескольких чисел строятся алгоритмы оптимального умножения. Пусть требуется перемножить n чисел a_i таких, что $|a_1|\le|a_2|\le\ldots\le|a_n|$. Погрешность произведения будет минимальной, если находить его по схеме: умножать a_1 последовательно на a_n, a_{n-1}, ... до тех пор, пока модуль частичного произведения не станет большим единицы, затем это частичное произведение умножать на a_2, a_3, ... до тех пор, пока новое частичное произведение не станет по модулю меньшим единицы, и так далее до исчерпывания всех сомножителей [42].

Естественно, что расплатой за выигрыш в точности при реализации таких алгоритмов будет проигрыш в скорости счета.

1.4. ПРИМЕРЫ НЕУСТОЙЧИВЫХ ЗАДАЧ И МЕТОДОВ

В силу неизбежного появления погрешностей в исходных данных задачи (в процессе создания математической модели изучаемого объекта или явления), а также погрешностей округления при ее решении, следует иметь представление о том, насколько чувствительными могут оказаться сами задачи и методы их решения к таким погрешностям. Рассмотрим несколько примеров проявления чрезмерной чувствительности.

Пример 1. Пусть требуется найти вещественное решение уравнения

$$(x-a)^n=\varepsilon,$$

где ε — очень малое положительное число, а натуральное n достаточно велико. Тогда естественно заменить ε нулем и положить $x\approx a$. Так как точное решение данного уравнения есть $x=a+\sqrt[n]{\varepsilon}$, то абсолютная погрешность при таком подходе составит величину $\sqrt[n]{\varepsilon}$. Много это или мало — судить об этом можно, придавая ε и n численные значения. Например, взяв $\varepsilon=10^{-10}$, $n=10$, получим абсолютную погрешность значения корня $x\approx a$,

равную 0.1. Относительная погрешность при этом может оказаться сколь угодно большой, если a взять сколь угодно малым.

Пример 2 (***пример Уилкинсона*** [3, 6, 75, 179 и др.]). Многочлен

$$P_{20}(x) := (x-1)(x-2)\ldots(x-20) \equiv x^{20} - 210x^{19} + \ldots + 20!$$

имеет 20 хорошо отделимых действительных корней

$$x_1 = 1, \quad x_2 = 2, \quad \ldots, \quad x_{20} = 20.$$

Предположим, что только в одном его коэффициенте, а именно, при x^{19} сделана ошибка порядка *macheps*: вместо -210 в развернутый вид многочлена $P_{20}(x)$ подставлено число $-\left(210 + 2^{-23}\right) \approx$ $\approx -\left(210 + 10^{-7}\right)$. Полученный при этом так называемый ***возмущенный многочлен*** будет иметь следующие корни (ограничимся записью трех цифр после запятой):

$$\begin{array}{lll} x_1 \approx 1.000, & x_6 \approx 6.000, & x_{12,13} \approx 11.794 \pm 1.652i, \\ x_2 \approx 2.000, & x_7 \approx 7.000, & x_{14,15} \approx 13.992 \pm 2.519i, \\ x_3 \approx 3.000, & x_8 \approx 8.007, & x_{16,17} \approx 16.731 \pm 2.813i, \\ x_4 \approx 4.000, & x_9 \approx 8.917, & x_{18,19} \approx 19.502 \pm 1.940i, \\ x_5 \approx 5.000, & x_{10,11} \approx 10.095 \pm 0.644i, & x_{20} \approx 20.847. \end{array}$$

Как видим, весьма малое возмущение (сопоставимое с точностью представления чисел в некоторых ЭВМ) всего лишь в одном коэффициенте даже качественно изменило набор корней данного многочлена: половина из них перестали быть действительными.

Пример 3. Линейная система

$$\begin{cases} x + 10y = 11, \\ 100x + 1001y = 1101 \end{cases}$$

имеет единственное решение $x = 1, y = -1$. Допустив абсолютную погрешность в 0.01 в правой части одного уравнения, получим ***возмущенную систему***

$$\begin{cases} x + 10y = 11.01, \\ 100x + 1001y = 1101 \end{cases}$$

с единственным решением $x = 11.01$, $y = 0$. Последнее никак не назовешь близким к решению исходной системы.

Пример 4. Для вычисления определенных интегралов вида

$$I_n := \frac{1}{e}\int_0^1 x^n e^x\,dx,$$

где $n \in \mathbf{N}$, с помощью метода интегрирования «по частям» легко вывести рекуррентную формулу

$$I_n = 1 - nI_{n-1}; \quad n = 1, 2, \ldots; \quad I_0 = 1 - \frac{1}{e}. \qquad (1.8)$$

В [10] приведена сводная таблица результатов подсчета I_n при $n = 0, 1, 2, \ldots, 14$ по формуле (1.8), полученных в 60-х годах на разных вычислительных машинах в Чехословакии, Восточной Германии, России. Эта таблица наглядно демонстрирует рост разброса значений интеграла при увеличении значения n. Не приводя здесь полностью этих результатов, отметим лишь, что при $n = 14$ спектр вычисленных значений интеграла I_{14} лежит в границах от -148 до 5356. Несмотря на то, что на некоторых ЭВМ вычисления велись с мантиссами, имеющими более десяти десятичных разрядов, ни один результат не имел ни одной верной цифры! Причина подобного явления — в численной неустойчивости схемы (1.8). Безупречная в теоретическом плане, т.е. с точки зрения аналитической математики, она совершенно непригодна с позиций вычислительной математики, поскольку неизбежная погрешность стартового значения I_0 при подсчете I_n увеличивается в $n!$ раз, т.е. катастрофически нарастает.

Если примеры 1–3 указывали прямо на существование ***неустойчивых задач****), то в четвертом примере видно яркое проявление ***неустойчивого метода*** вычислений. Последним, как правило, есть альтернативы. Например, для вычисления рассматриваемого там интеграла I_n при конкретном значении n можно применять квадратурные формулы, т.е. формулы, специально приспособленные для приближенного вычисления определенных интегралов. А можно воспользоваться и равенством (1.8), только переписать его в виде

$$I_{n-1} = \frac{1}{n}(1 - I_n). \qquad (1.9)$$

---

*) См. также пример неустойчивой задачи на собственные значения для $n \times n$-матрицы в § 4.5.

Учитывая, что $0 < I_{n+1} < I_n$ и $I_n \xrightarrow[n\to\infty]{} 0$, для подсчета I_n при некотором фиксированном $n = k$ (и для меньших этого k индексов) можно задаться значением $I_N := 0$ и вести счет по формуле (1.9) при $n = N,\ N-1,\ \ldots,\ k+1$. Так как начальная погрешность на каждом шаге теперь уменьшается в n раз, то такой алгоритм будет ***численно устойчивым***. Значение N при этом может быть определено теоретически или экспериментально [6, 10].

Более подробно тема устойчивости и неустойчивости задач и методов развивается при изучении численных процессов решения дифференциальных уравнений (см. гл. 16, а также §§ 2.6, 13.3, 17.3, 18.1, 20.2). Для решения задач, требующих особого учета возмущений и обобщения самого понятия решения, в последние несколько десятилетий разрабатываются особые подходы, приводящие к построению специальных устойчивых численных методов, называемых методами регуляризации. Первые представления о таких методах можно получить из § 1.7.

1.5. ОБУСЛОВЛЕННОСТЬ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ СИСТЕМ

Учитывая распространенность систем линейных алгебраических уравнений (ибо часто именно к ним сводится на определенном этапе процесс математического моделирования), попытаемся количественно охарактеризовать степень неопределенности этих задач. Знание таких характеристик позволяет обоснованно судить о корректности моделей, грамотно подбирать методы и строить алгоритмы, правильно трактовать полученные результаты.

Рассмотрим линейную алгебраическую систему, записанную в виде векторно-матричного уравнения

$$\mathbf{Ax} = \mathbf{b}, \tag{1.10}$$

где $\mathbf{A}$ — невырожденная $n \times n$-матрица коэффициентов данной системы; $\mathbf{b}$ — ненулевой n-мерный вектор свободных членов; $\mathbf{x}$ — n-мерный вектор неизвестных (решение, если трактовать (1.10) как верное равенство).

Пусть правая часть (1.10) получила приращение («возмущение») $\Delta\mathbf{b}$, т.е. вместо истинного вектора $\mathbf{b}$ используется приближенный вектор $\mathbf{b} + \Delta\mathbf{b}$. Реакцией решения $\mathbf{x}$ на возмущение $\Delta\mathbf{b}$ правой части будет вектор поправок $\Delta\mathbf{x}$, т.е. если $\mathbf{x}$ — решение (1.10), то $\mathbf{x} + \Delta\mathbf{x}$ — решение уравнения

$$\mathbf{A}(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}. \tag{1.11}$$

Понимая под абсолютной погрешностью приближенного вектора норму*) разности между точным и приближенным векторами, а под относительной погрешностью — отношение абсолютной погрешности к норме вектора (точного или приближенного), выясним связь между относительными погрешностями вектора свободных членов и вектора-решения. Иначе, получим оценку вида

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \le (?)\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|},$$

где $\|\bullet\|$ — какая-либо векторная норма, а (?) — неизвестный пока коэффициент связи.

Подставляя (1.10) в (1.11), видим, что поправка $\Delta\mathbf{x}$ связана с возмущением $\Delta\mathbf{b}$ таким же, как и (1.10), равенством

$$\mathbf{A}\Delta\mathbf{x} = \Delta\mathbf{b},$$

из которого находим ее явное выражение

$$\Delta\mathbf{x} = \mathbf{A}^{-1}\Delta\mathbf{b}. \tag{1.12}$$

Нормируя равенства (1.10) и (1.12), будем иметь

$$\|\mathbf{b}\| \le \|\mathbf{A}\| \cdot \|\mathbf{x}\| \quad \text{и} \quad \|\Delta\mathbf{x}\| \le \left\|\mathbf{A}^{-1}\right\| \cdot \|\Delta\mathbf{b}\|,$$

где матричная норма должна быть согласованной с выбранной векторной нормой. Эти два числовых неравенства одинакового смысла можно перемножить:

$$\|\mathbf{b}\| \cdot \|\Delta\mathbf{x}\| \le \|\mathbf{A}\| \cdot \|\mathbf{x}\| \cdot \left\|\mathbf{A}^{-1}\right\| \cdot \|\Delta\mathbf{b}\|.$$

Из последнего делением на $\|\mathbf{b}\| \cdot \|\mathbf{x}\|$ получаем искомую связь

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \|\mathbf{A}\| \cdot \left\|\mathbf{A}^{-1}\right\| \cdot \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}. \tag{1.13}$$

Положительное число $\|\mathbf{A}\| \cdot \left\|\mathbf{A}^{-1}\right\|$ — коэффициент этой связи — называют ***числом (мерой) обусловленности*** матрицы $\mathbf{A}$ и обозначают $\operatorname{cond}\mathbf{A}$ (от английского слова *conditioned* — «обусловленный»)**). Распространены также обозначения $\nu(\mathbf{A})$, $\chi(\mathbf{A})$, $\mu(\mathbf{A})$.

---

*) См. приложение 1.

**) Имеется более общее определение числа обусловленности [194]:

$$\operatorname{cond}\mathbf{A} := \sup_{\mathbf{x}\ne 0}\frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} \Big/ \inf_{\mathbf{y}\ne 0}\frac{\|\mathbf{Ay}\|}{\|\mathbf{y}\|}.$$

Здесь $\operatorname{cond}\mathbf{A}$ задается через векторные нормы и может быть применено к вырожденным матрицам $\mathbf{A}$. В случае обратимых матриц $\mathbf{A}$ при использовании согласованных матричных норм отсюда получается $\operatorname{cond}\mathbf{A} = \|\mathbf{A}\| \cdot \|\mathbf{A}\|^{-1}$.

Легко показать, что то же самое число $\operatorname{cond}\mathbf{A} = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|$ служит коэффициентом роста относительных погрешностей при неточном задании элементов матрицы $\mathbf{A}$ в (1.10). А именно, если матрица $\mathbf{A}$ получила возмущение $\Delta\mathbf{A}$ и $\mathbf{x}+\Delta\mathbf{x}$ — решение возмущенной системы

$$(\mathbf{A}+\Delta\mathbf{A})(\mathbf{x}+\Delta\mathbf{x}) = \mathbf{b},$$

то справедливы неравенства

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \operatorname{cond}\mathbf{A} \cdot \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}+\Delta\mathbf{A}\|} \quad \text{и} \quad \frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}+\Delta\mathbf{x}\|} \le \operatorname{cond}\mathbf{A} \cdot \frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}. \qquad (1.14)$$

Замечание 1.3. Можно получить оценки, обобщающие в определенном смысле (1.13) и (1.14) в случае одновременного возмущения матрицы $\mathbf{A}$ системы (1.10) и ее правой части $\mathbf{b}$. Более того, величина $\nu(\mathbf{A}) := \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|$ является также мерой обусловленности линейного обратимого оператора $\mathbf{A}$ в произвольном нормированном пространстве. Справедливо следующее утверждение [177].

Теорема 1.1. *Пусть* $\mathbf{A}\mathbf{x} = \mathbf{b}$ *— данное, а* $\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$ *— возмущенное линейные операторные уравнения с относительными уровнями возмущений* $\delta_{\mathbf{A}} \ge \frac{\|\mathbf{A}-\tilde{\mathbf{A}}\|}{\|\mathbf{A}\|}$ *и* $\delta_{\mathbf{b}} \ge \frac{\|\mathbf{b}-\tilde{\mathbf{b}}\|}{\|\mathbf{b}\|}$. *Тогда, если* $\delta_{\mathbf{A}}\nu(\mathbf{A}) < 1$, *то эти уравнения одновременно однозначно разрешимы и справедлива оценка относительной погрешности решения, имеющая вид*

$$\frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \le \frac{\nu(\mathbf{A})}{1-\delta_{\mathbf{A}}\cdot\nu(\mathbf{A})}\cdot\delta_{\mathbf{b}} + \frac{\nu^2(\mathbf{A})}{1-\delta_{\mathbf{A}}\cdot\nu(\mathbf{A})}\cdot\delta_{\mathbf{A}}.$$

Итак, неравенства (1.13) и (1.14) показывают, что *чем больше число обусловленности, тем сильнее сказывается на решении линейной системы ошибка в исходных данных*. Грубо говоря, если $\operatorname{cond}\mathbf{A} = O(10^p)$ и исходные данные имеют погрешность в l-м знаке после запятой, то независимо от способа решения системы (1.10) в результате можно гарантировать не более $l-p$ знаков после запятой.

Если число $\operatorname{cond}\mathbf{A}$ велико, то система считается плохо обусловленной. Говорить о том, «что такое хорошо и что такое плохо», в отрыве от контекста решаемой задачи почти бессмысленно, так как здесь может играть роль размерность задачи, точность, с которой должно быть найдено ее решение, точность

представления чисел в ЭВМ и т.п. Однако можно дать оценку снизу числа обусловленности. А именно, если используются подчиненные матричные нормы (для которых норма единичной матрицы есть единица), то, очевидно,

$$\operatorname{cond}\mathbf{A} = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| \ge \|\mathbf{A}\cdot\mathbf{A}^{-1}\| = \|\mathbf{E}\| = 1,$$

т.е. число обусловленности не может быть меньше 1. Можно также указать верхнюю границу для чисел обусловленности, превышение которой при решении линейных систем на конкретной ЭВМ может привести к заведомо ложным результатам. Так, решение считается ненадежным, если $\operatorname{cond}\mathbf{A} \ge (macheps)^{-1}$ или даже $\operatorname{cond}\mathbf{A} \ge (macheps)^{-0.5}$ [68]. При этом заметим, что масштабированием матрицы $\mathbf{A}$ путем умножения на скаляр α ее обусловленность не улучшить, ибо

$$\operatorname{cond}(\alpha\mathbf{A}) = \|\alpha\mathbf{A}\| \cdot \|(\alpha\mathbf{A})^{-1}\| = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| = \operatorname{cond}\mathbf{A}.$$

Приемы более-менее эффективного масштабирования линейных систем вида (1.10), повышающего устойчивость их решений, можно найти, например, в [71].

Классическим примером плохо обусловленной матрицы является так называемая ***матрица Гильберта****)

$$\mathbf{H}_n = \left(\frac{1}{i+j-1}\right)_{i,j=1}^{n},$$

демонстрирующая катастрофическое возрастание числа обусловленности с ростом размерности [138, 183]. Так, уже при $n=8$ $\operatorname{cond}\mathbf{H}_8 > 10^{10}$ и обратная матрица $\mathbf{H}_8^{-1}$, полученная на машине с точностью представления чисел порядка 10^{-8}, может не содержать ни одного верного знака.

Очевидно, число обусловленности зависит от выбора матричной нормы (индуцированной, как правило, той или иной векторной нормой, в терминах которой характеризуется относительная погрешность решения алгебраической системы). Однако

*) Эта матрица — не просто плод воображения. Ее появление закономерно, например, при нахождении коэффициентов многочленов наилучшего среднеквадратического приближения методом наименьших квадратов (§ 10.3).

нетрудно получить оценку числа обусловленности через собственные числа матрицы. Действительно, пусть собственные числа λ_i матрицы **A** упорядочены по модулю:

$$|\lambda_1| \ge |\lambda_2| \ge \ldots \ge |\lambda_n|,$$

т.е. ***спектральный радиус*** матрицы **A** есть $\rho_{\mathbf{A}} = |\lambda_1|$. Тогда в силу известного неравенства $\rho_{\mathbf{A}} \le \|\mathbf{A}\|$ и соотношения между собственными числами прямой и обратной матриц, имеем

$$\|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| \ge \rho_{\mathbf{A}} \cdot \rho_{\mathbf{A}^{-1}} = |\lambda_1| \cdot \frac{1}{|\lambda_n|}.$$

Таким образом, оценкой снизу меры обусловленности матрицы **A** может служить величина $\left|\frac{\lambda_1}{\lambda_n}\right|$ (называемая иногда ***числом обусловленности Тодда*** [РЖМат, 1983, 10Б937] или ***Тодта*** [126]). Для симметричных матриц эта оценка и на самом деле является числом обусловленности, соответствующим спектральной норме матрицы (индуцированной евклидовой нормой вектора). Учитывая смысл собственных чисел матрицы, можно сказать, что число обусловленности показывает величину отношения наибольшего коэффициента растяжения вектора посредством линейного преобра–зования **A** к наименьшему.

Следует отметить, что непосредственный подсчет чисел обусловленности, особенно при большой размерности матриц, является весьма дорогостоящим делом из-за необходимости обращать матрицы или находить их собственные значения. Поэтому зачастую о приемлемости порядка возможного роста относительной погрешности результата решения какой-либо алгебраической задачи относительно данной матрицы судят либо по каким-то достаточным признакам (например, по доминированию элементов главной диагонали матрицы), либо на основе теоретического изучения матрицы [138, 152], либо путем применения специальных алгоритмов приближенного оценивания cond **A** [68, 127, 184]. Исследование матриц на обусловленность может быть естественным образом связано со способом решения той или иной алгебраической задачи относительно данной матрицы.

Прокомментируем теперь пример неустойчивой системы, приведенной в примере 3 § 1.4.

Матрица коэффициентов системы $\mathbf{A} = \begin{pmatrix} 1 & 10 \\ 100 & 1001 \end{pmatrix}$ имеет обратную $\mathbf{A}^{-1} = \begin{pmatrix} 1001 & -10 \\ -100 & 1 \end{pmatrix}$. Следовательно, число обусловленности в матричной норме, индуцированной векторной нормой-максимум (иначе, нормой l_∞), есть

$$\nu_\infty(\mathbf{A}) = 1101 \cdot 1011 = 1113111 > 10^6.$$

Учитывая, что в данном примере $\mathbf{b} = \begin{pmatrix} 11 \\ 1101 \end{pmatrix}$, $\Delta\mathbf{b} = \begin{pmatrix} 0.01 \\ 0 \end{pmatrix}$, на основе (1.13) получаем оценку относительной погрешности решения в l_∞-нормах

$$\delta_{\mathbf{x}} \le \nu_\infty(\mathbf{A}) \cdot \delta_{\mathbf{b}} = 1113111 \cdot \frac{0.01}{1101} = 10.11.$$

Так как норма-максимум решения $\mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ равна 1, оценка абсолютной погрешности $\|\Delta\mathbf{x}\|$ решения равна

$$\|\Delta\mathbf{x}\| = \|\mathbf{x}\| \cdot \delta_{\mathbf{x}} \le 10.11.$$

Как видим, решение $\mathbf{x} + \Delta\mathbf{x} = \begin{pmatrix} 11.01 \\ 0 \end{pmatrix}$ возмущенной системы вписывается в оценку

$$\|\mathbf{x} + \Delta\mathbf{x}\| \le \|\mathbf{x}\| + \|\Delta\mathbf{x}\| \le 1 + 10.11 = 11.11.$$

Аналогичный результат может быть получен через число обусловленности Тодда. Решая характеристическое уравнение

$$\lambda^2 - 1002\lambda + 1 = 0,$$

находим собственные числа матрицы **A**: $\lambda_1 \approx 1002$ и $\lambda_2 \approx 0.000998$, дающие оценку

$$\mathrm{cond}\,\mathbf{A} \ge \frac{\lambda_1}{\lambda_2} \approx 1004000 > 10^6.$$

На данном примере также можно наглядно убедиться в том, что *малость* ***невязки*** $\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}$ *плохо обусловленной системы еще не говорит о близости приближенного решения* $\tilde{\mathbf{x}}$ *к точно-*

му $\mathbf{x}$. Действительно, невязки $\mathbf{r}_1$ и $\mathbf{r}_2$ векторов $\tilde{\mathbf{x}}_1 = \begin{pmatrix} 11.01 \\ 0 \end{pmatrix}$ и $\tilde{\mathbf{x}}_2 = \begin{pmatrix} 1 \\ 1.1 \end{pmatrix}$ для основной (невозмущенной) системы из примера 3 § 1.4 имеют нормы соответственно $\|\mathbf{r}_1\|_\infty = 0.01$ и $\|\mathbf{r}_2\|_\infty = 100.1$. Вектор $\tilde{\mathbf{x}}_2$, явно более близкий к точному решению $\mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, чем $\tilde{\mathbf{x}}_1$, имеет существенно большую невязку! Объяснение этому парадоксальному, на первый взгляд, факту можно найти в книге [26].

Двумерный случай допускает простую геометрическую трактовку понятия обусловленности. Плохая обусловленность системы двух уравнений с двумя неизвестными означает, что прямые, являющиеся геометрическими образами уравнений, пересекаются на координатной плоскости под очень острым углом. В этом случае небольшое искажение в данных, интерпретируемое как параллельный перенос (при возмущении свободного члена) или поворот прямых (при возмущении матрицы коэффициентов), приводит к значительному перемещению их точки пересечения, т.е. геометрического образа решения.

1.6. ПОГРЕШНОСТИ КОРНЕЙ СКАЛЯРНЫХ УРАВНЕНИЙ С ПРИБЛИЖЕННЫМИ КОЭФФИЦИЕНТАМИ

Пусть решается алгебраическое или трансцендентное уравнение вида

$$f(x) = 0. \tag{1.15}$$

Это уравнение, являясь нелинейным, может иметь один или несколько корней или не иметь их вовсе. Каждый корень так или иначе зависит от числовых данных уравнения (1.15). Эти числовые данные (будем иногда называть их коэффициентами уравнения) можно считать фиксированными значениями параметров, т.е. уравнение (1.15) целесообразно рассматривать в качестве представителя семейства уравнений

$$f(x, a_1, a_2, \ldots, a_m) = 0 \tag{1.16}$$

и наличие корней у (1.15) связывать с существованием неявной функции $x = \varphi(a_1, a_2, \ldots, a_m)$ при данном наборе значений параметров (коэффициентов) $a_1, a_2, \ldots, a_m$. Поскольку эти числа, как

правило, точно не известны (грубость модели, неточность измерений, усечение чисел при вводе в ЭВМ и т.п.), встает вопрос о том, как влияет погрешность коэффициентов уравнения (1.15) на погрешности его корней. Иначе, с какой точностью имеет смысл решать данное уравнение, если известно, что его коэффициенты не точны, но имеется информация об уровне их погрешностей?

Реально можно оценить лишь главную часть погрешности корня, понимая под ней, как и в § 1.1, модуль дифференциала. В данном случае речь идет о дифференциале неявной функции нескольких переменных. Подходящую основу для этого находим в математическом анализе: если, например, выполняются условия, при которых уравнение $F(x, y, z) = 0$ неявно определяет дифференцируемую функцию $z = z(x, y)$, то выражение дифференциала dz этой функции можно получить из равенства $dF = 0$, а именно,

$$dz = -\frac{1}{F'_z}\left(F'_x\,dx + F'_y\,dy\right).$$

Аналогично, если $x = \varphi(a_1, a_2, \ldots, a_m)$ — корень уравнения (1.16), то линейная часть его изменения, соответствующая изменениям аргументов — коэффициентов $a_1, a_2, \ldots, a_m$, равна

$$dx = -\frac{1}{\dfrac{\partial f}{\partial x}}\left(\frac{\partial f}{\partial a_1}d\,a_1 + \frac{\partial f}{\partial a_2}d\,a_2 + \ldots + \frac{\partial f}{\partial a_m}d\,a_m\right).$$

Переходя здесь к модулям и заменяя истинные абсолютные погрешности коэффициентов $\Delta a_i (= |da_i|)$ их оценками Δ_{a_i}, получим ***формулу для оценки абсолютной погрешности корня***:

$$|\Delta x| \lesssim \frac{1}{\left|\dfrac{\partial f}{\partial x}\right|}\left(\left|\frac{\partial f}{\partial a_1}\right|\Delta_{a_1} + \left|\frac{\partial f}{\partial a_2}\right|\Delta_{a_2} + \ldots + \left|\frac{\partial f}{\partial a_m}\right|\Delta_{a_m}\right). \tag{1.17}$$

Заметим, что при вычислении значений частных производных в (1.17) следует пользоваться фиксированными значениями коэффициентов $a_1, a_2, \ldots, a_m$ (такими, какими они используются при нахождении корня) и приближенным значением того корня, степень неопределенности которого устанавливается. Ясно, что для разных корней одного и того же уравнения значения этой величины могут сильно различаться.

Следуя [72], получаемую с помощью (1.17) оценку будем называть ***безусловной абсолютной погрешностью*** приближенного корня x уравнения (1.16) с приближенными коэффициентами и обозначать б.п. x. Это же x есть точный корень уравнения

(1.15), в то время как при численном решении уравнения вместо x будет получено некоторое приближение к нему $\bar{\bar{x}}$. Возникающую при этом остаточную погрешность или погрешность метода назовем ***условной погрешностью*** и обозначим у.п. x.

Итак, в ходе численного решения уравнения (1.15) в предположении, что его коэффициенты точны, находится приближенный корень $\bar{\bar{x}}$ с условной погрешностью у.п. x. Это означает, что точный корень x уравнения (1.15) при данном предположении лежит на интервале ($\bar{\bar{x}}$ – у.п.x, $\bar{\bar{x}}$ + у.п.x). Зная погрешности коэффициентов и приближенный корень $\bar{\bar{x}}$, можно подсчитать безусловную погрешность б.п.x. Истинная величина корня x^* уравнения с приближенными коэффициентами — это потенциально любое число из интервала (x – б.п.x, x + б.п.x). Ясно, что гипотетическое значение x^* — решение рассчитываемой модели — может отличаться от реально полученного значения $\bar{\bar{x}}$ на величину у.п.x + б.п.x, т.е. *полная погрешность корня уравнения с приближенными коэффициентами складывается из погрешностей условной и безусловной*. Так как величиной условной погрешности распоряжается вычислитель, причем ее уменьшение в разумных пределах не вызывает, как правило, больших затруднений, то обычно, чтобы не увеличивать полную погрешность корня приближенного уравнения, условную погрешность задают неравенством

$$\text{у.п.}x \le \text{б.п.}x \tag{1.18}$$

или даже

$$\text{у.п.}x \le 0.1\ \text{б.п.}x.$$

Если воспользоваться ***правилом Ньютона****) оценки близости приближенного корня (нуля) $\bar{\bar{x}}$ дифференцируемой функции $f(x)$ к точному корню x

$$\left|\bar{\bar{x}} - x\right| \lessapprox \left|\frac{f(\bar{\bar{x}})}{f'(\bar{\bar{x}})}\right| \tag{1.19}$$

---

*) По формуле конечных приращений Лагранжа при некотором $\Theta \in (\bar{\bar{x}}, x)$ (или $(x, \bar{\bar{x}})$) для дифференцируемой функции $f(x)$ выполняется равенство

$$f(\bar{\bar{x}}) - f(x) = f'(\Theta)(\bar{\bar{x}} - x).$$

Предполагая, что x — корень, $\bar{\bar{x}} \approx x$, а значит и $\Theta \approx \bar{\bar{x}}$, имеем

$$f(\bar{\bar{x}}) \approx f'(\bar{\bar{x}})(\bar{\bar{x}} - x),$$

откуда следует правило Ньютона (1.19).

и оценкой (1.17), то можно на основе (1.18) получить простой критерий окончания процесса численного решения скалярного уравнения (1.15) с приближенными коэффициентами:

уточнение корня $\bar{\bar{x}} \approx x$ ведется до тех пор, пока не выполнится условие

$$\left|f(\bar{\bar{x}})\right| \le \left|\frac{\partial f}{\partial a_1}\right|\Delta_{a_1} + \left|\frac{\partial f}{\partial a_2}\right|\Delta_{a_2} + \ldots + \left|\frac{\partial f}{\partial a_m}\right|\Delta_{a_m}. \tag{1.20}$$

Как видим, бóльшую роль в неравенстве (1.20) играют модули частных производных данной функции по параметрам, называемые ***коэффициентами чувствительности***.

В случае, когда уравнение (1.15) — алгебраическое, т.е.

$$f(x) \equiv P(x) := a_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1}x + a_n, \tag{1.21}$$

его корни — функции коэффициентов многочлена $a_0, a_1, \ldots, a_n$, а коэффициенты чувствительности суть числа $\left|x^n\right|, \left|x^{n-1}\right|, \ldots, |x|, 1$, получающиеся при подстановке сюда вместо x приближенных корней многочлена. Неравенство типа (1.20), закладываемое в процесс уточнения корня $\bar{\bar{x}}$ многочлена (1.21) с приближенными коэффициентами $a_0(\pm\Delta_{a_0})$, $a_1(\pm\Delta_{a_1})$, … , $a_n(\pm\Delta_{a_n})$, имеет вид

$$\left|P(\bar{\bar{x}})\right| \le \left|\bar{\bar{x}}^n\right|\Delta_{a_0} + \left|\bar{\bar{x}}^{n-1}\right|\Delta_{a_1} + \ldots + \left|\bar{\bar{x}}\right|\Delta_{a_{n-1}} + \Delta_{a_n}.$$

Легко заметить, какую роль играют погрешности тех или иных коэффициентов в зависимости от того, малы или велики модули корней.

Получим, наконец, связь между относительными погрешностями коэффициентов и корней многочлена.

На основе оценки (1.17) для многочлена (1.21) имеем

$$\delta x = \left|\frac{\Delta x}{x}\right| \lessapprox \frac{\sum_{i=0}^{n}\left|\frac{\partial P}{\partial a_i}\right|\Delta_{a_i}}{|x|\cdot\left|\frac{\partial P}{\partial x}\right|} = \frac{\sum_{i=0}^{n}\left|x^{n-i}\right|\cdot|a_i|\cdot\delta_{a_i}}{|x|\cdot\left|\sum_{i=0}^{n-1}(n-i)a_i x^{n-1-i}\right|} = \frac{\sum_{i=0}^{n}\left|a_i x^{n-i}\right|\delta_{a_i}}{\left|\sum_{i=0}^{n-1}(n-i)a_i x^{n-i}\right|}.$$

Если сделать допущение, что все коэффициенты многочлена имеют одинаковый уровень δ относительных погрешностей, или

положить $\max\limits_{i=0,\ldots,n} \delta_{a_i} = \delta$, то для граничной относительной погрешности δ_x простого ненулевого корня x будем иметь приближенную оценку

$$\delta_x \lesssim \frac{\sum\limits_{i=0}^{n} \left| a_i x^{n-i} \right|}{\left| \sum\limits_{i=0}^{n-1} (n-i) a_i x^{n-i} \right|} \cdot \delta. \qquad (1.22)$$

Коэффициент

$$\nu_x := \frac{\sum\limits_{i=0}^{n} \left| a_i x^{n-i} \right|}{\left| \sum\limits_{i=0}^{n-1} (n-i) a_i x^{n-i} \right|}$$

связи (1.22) относительных погрешностей корней и коэффициентов (для каждого корня свой) по аналогии с терминологией предыдущего параграфа естественно назвать ***числом*** или ***мерой обусловленности ненулевого простого корня*** *) ***многочлена с приближенными коэффициентами***.

Замечание 1.4. В книге [75] мерой обусловленности простого корня $x = x(a_1, a_2, \ldots, a_n)$ многочлена (1.21) с $a_0 = 1$ считается норма градиента $\left(\frac{\partial x}{\partial a_1}; \frac{\partial x}{\partial a_2}; \ldots; \frac{\partial x}{\partial a_n} \right)$. Если все корни $x_1, x_2, \ldots, x_n$ — простые, то обусловленность многочлена характеризуется нормой матрицы Якоби $\left(\frac{\partial x_i}{\partial a_j} \right)_{i,\, j=1}^{n}$.

Обращаясь к примеру Уилкинсона в § 1.4, с помощью подсчета чисел обусловленности корней можно грубо объяснить картину разного влияния на разные корни внесенного в один коэффициент малого возмущения. Если для первого корня число обусловленности $\nu_x \approx 400$, то с увеличением номера корня это число значительно возрастает, достигая, например, на пятнадцатом корне величины большей, чем 10^{10}. Последний (двадцатый) корень имеет меньшее, чем предыдущий, число обусловленности, сравнимое с величиной $1/macheps$ (значения несколько иначе определенных чисел обусловленности можно найти в [3]).

*) В случае нулевого или кратного корня x знаменатель дроби обращается в нуль.

Отметим, что в числителе выражения для подсчета числа обусловленности ν_x могут отсутствовать слагаемые, соответствующие теоретически точным и при этом точно реализуемым при вводе в ЭВМ значениям коэффициентов многочлена. Это видно из оценки δ_x, где некоторые из Δ_{a_i} могут быть тождественно равными нулю (т.е. соответствующие коэффициенты не считаются параметрами приближенного уравнения (1.16)).

1.7. КОРРЕКТНЫЕ И НЕКОРРЕКТНЫЕ ЗАДАЧИ. ПОНЯТИЕ О МЕТОДАХ РЕГУЛЯРИЗАЦИИ

В начале XX века при выяснении вопроса о соответствии математических и физических моделей задач естествознания впервые было введено понятие корректности математической задачи. В достаточно общем виде для операторного уравнения

$$Ay = f, \qquad (1.23)$$

где $A: Y \to F$ — некоторый оператор, действующий из метрического пространства Y в метрическое пространство F (обычно Y и F — полные метрические пространства, в частности, банаховы), корректность вводится следующим образом.

Определение 1.1. *Задача нахождения элемента* $y \in Y$ *по заданному элементу* $f \in F$ *из уравнения* (1.23) *называется* ***корректно поставленной по Адамару*** *) *для тройки* (Y, A, F) *или просто* ***корректной***, *если:*

1) *при каждом* $f \in F$ *существует решение* $y \in Y$;

2) *это решение* y *единственно в* Y;

3) *решение непрерывно зависит от правой части уравнения* (1.23) (*т.е. из сходимости* $f_n \to f$ *по метрике пространства* F *следует сходимость* $y_n \to y$ *по метрике пространства* Y).

При нарушении любого из этих трех условий задача (1.23) *считается* ***некорректной***.

Первые два условия корректности, фигурирующие в опре-

*) Адамáр Жак Саломон (1865–1963) — французский математик. Широко известен своими исследованиями по теории чисел, теории целых аналитических функций, теории дифференциальных уравнений, функциональному анализу, механике.

делении 1.1, означают существование обратного оператора $A^{-1}: F \to Y$, третье условие — его непрерывность на всем пространстве F.

Часто требование непрерывной зависимости решения от правой части уравнения (1.23) заменяют условием его устойчивости:

$$\forall \varepsilon > 0 \quad \exists \delta := \delta(\varepsilon): \quad \rho_F(f_1, f_2) < \delta \Rightarrow \rho_Y(y_1, y_2) < \varepsilon,$$

где f_1, f_2 — произвольные элементы F, а $y_1, y_2 \in Y$ — отвечающие им решения ($\rho_F(\cdot,\cdot)$ и $\rho_Y(\cdot,\cdot)$ — метрики пространств F и Y соответственно).

Несколько более широкая трактовка третьего условия корректности состоит в том, что *в корректно поставленных задачах небольшие изменения в исходных данных* (малые возмущения f и A в (1.23)) *вызывают небольшие изменения решения.*

До середины XX века математики не занимались теорией некорректных по Адамару задач, поскольку считалось, что эти задачи не имеют физического смысла. Однако содержательных некорректных задач, требующих математического обоснования и создания устойчивых методов их решения, в прикладных областях накопилось значительное множество. В большинстве своем - это задачи, связанные с созданием систем автоматической математической обработки результатов наблюдений и физических экспериментов. Таковыми являются задачи восстановления сигнала, обратные задачи теплопроводности, геофизики, астрофизики и ряд других. С развитием науки и техники (в частности, цифровой) необходимость в умении решать некорректные задачи все возрастает.

Рассмотрим простейший **пример**, демонстрирующий нарушение тех или иных условий корректности и позволяющий понять, как можно избавиться от первых двух условий, т.е. сузить класс некорректных по Адамару задач.

Запишем три двумерные системы линейных уравнений:

$$\begin{cases} x_1 + x_2 = 1, \\ x_1 + x_2 = 1 \end{cases} \quad \text{(система А);}$$

$$\begin{cases} x_1 + x_2 = 1, \\ x_1 + 1.00001 x_2 = 1.000005 \end{cases} \quad \text{(система Б);}$$

$$\begin{cases} x_1 + x_2 = 0.5, \\ x_1 + x_2 = 1.5 \end{cases} \quad \text{(система В);}$$

Первая из них имеет решение, но не единственное: $\mathbf{x}^* = (x_1; 1 - x_1)^T$, где x_1 — свободная переменная. Здесь нарушено второе условие корректности. В этом случае выход находят в том, что среди бесчисленного множества решений ищется решение, ближайшее к заданной точке $\mathbf{x}^0 = (x_1^0; x_2^0)^T$, называемой ***пробным решением***. Такая точка может быть известна из каких-то априорных соображений (например, из физического смысла задачи); в противном случае полагают $\mathbf{x}^0 = \mathbf{0}$.

Итак, если требуется найти ближайшее к $\mathbf{x}^0 = \mathbf{0}$ решение системы А, нужно решить задачу минимизации

$$\rho_{\mathbf{R}_2}\left(\begin{pmatrix} x_1 \\ 1 - x_1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right) \to \min.$$

Взяв метрику евклидова пространства $\mathbf{R}_2$, имеем

$$x_1^2 + (1 - x_1)^2 \to \min \quad \Leftrightarrow \quad 2x_1 - 2(1 - x_1) = 0,$$

откуда получаем $x_1 = 0.5$, $1 - x_1 = 0.5$, т.е. за решение системы А принимаем $\mathbf{x}_0^* = (0.5;\ 0.5)^T$.

Возвращаясь на время к общей задаче (1.23), дадим следующее

Определение 1.2. *Решение уравнения* (1.23), *ближайшее к заданному пробному решению* y^0, *называется* ***нормальным относительно*** y^0 ***решением*** (*или просто* ***нормальным решением***, *если* $y^0 = 0$).

Другими словами, нормальное относительно y^0 решение уравнения (1.23) — это проекция пробного решения y^0 на множество решений этого уравнения. Если указанное множество выпукло, то как известно, такая проекция, а следовательно, и нормальное решение единственно.

Обратимся теперь к системе Б. Легко убедиться, что решение этой системы существует и единственно: $\mathbf{x}_0^* = (0.5;\ 0.5)^T$. Произведем небольшие манипуляции с числами в этой системе. Если в ее правой части отбросить малое возмущение $\varepsilon = 0.000005$, то полученная в результате этого система

$$\begin{cases} x_1 + x_2 = 1, \\ x_1 + 1.00001 x_2 = 1 \end{cases}$$

будет иметь решение $(1; 0)^T$, значительно отличающееся от решения $(0.5;\ 0.5)^T$ системы Б. Отбросив и малую добавку 0.00001

в коэффициенте при x_2 придем к системе

$$\begin{cases} x_1 + x_2 = 1, \\ x_1 + x_2 = 1.000005, \end{cases}$$

вовсе не имеющей решений. Как видим, малые изменения в исходных данных системы Б влекут не только большие количественные изменения решения, но и меняют ситуацию с его существованием и единственностью, т.е. имеет место нарушение третьего требования в определении 1.1 корректности при выполненных первых двух для исходной задачи. Не вызывает сомнений факт, что для линейных алгебраических систем неустойчивость напрямую связана с их плохой обусловленностью (убедитесь, что число обусловленности Тодда (§ 1.5) для матрицы системы Б приближенно равно 50000).

Система В явно противоречива. Казалось бы, на этом следует завершить разговор об этой системе: нет решения и нет. Однако, нередко случается, что несовместные системы отражают реальные ситуации, в которых решение в каком-то смысле должно быть. Встает вопрос о расширении понятия решения линейной системы.

Еще в начале XIX века Гаусс и Лежандр*) независимо друг от друга предложили решать переопределенные, как правило, несовместные системы методом наименьших квадратов (более подробно об этом см. в § 10.1). Суть их предложения состоит в том, что за решение таких систем вида $\mathbf{Ay} = \mathbf{f}$ принимается вектор $\mathbf{y}$, минимизирующий функцию

$$\Phi(\mathbf{y}) := \|\mathbf{Ay} - \mathbf{f}\|_2^2 = (\mathbf{Ay} - \mathbf{f},\ \mathbf{Ay} - \mathbf{f}).$$

Эта функция всегда имеет неотрицательный минимум, который можно эффективно найти ввиду отсутствия локальных минимумов, отличных от глобального.

Определение 1.3. *Вектор y, на котором достигается* $\min \Phi(\mathbf{y})$, *называется* ***псевдорешением*** *системы* $\mathbf{Ay} = \mathbf{f}$.

*) Га́усс Карл Фридрих (1777–1855) — известнейший немецкий математик. Его труды оказали глубокое влияние на развитие алгебры, теории чисел, дифференциальной геометрии, теории тяготения, геодезии, теории электричества и магнетизма, астрономии и др.

Лежа́ндр Адриен Мари (1752–1833) — французский математик. Известен своими работами по математическому анализу, вариационному исчислению, теории чисел. Обосновал и развил теорию геодезических измерений.

Из вида $\Phi(\mathbf{y})$ ясно, что если $\min \Phi(\mathbf{y}) = 0$, то $\mathbf{y}^* := \arg\min \Phi(\mathbf{y})$ есть точное решение системы $\mathbf{Ay} = \mathbf{f}$, т.е. псевдорешение обобщает понятие решения в привычном понимании.

Для нахождения псевдорешения системы $\mathbf{Ay} = \mathbf{f}$ вместо решения экстремальной задачи $\Phi(\mathbf{y}) \to \min$ можно решать симметричную систему

$$\mathbf{A}^T \mathbf{Ay} = \mathbf{A}^T \mathbf{f}. \tag{1.24}$$

Действительно, так как

$$\Phi(\mathbf{y}) = (\mathbf{Ay} - \mathbf{f},\ \mathbf{Ay} - \mathbf{f}) = (\mathbf{Ay},\ \mathbf{Ay}) - 2(\mathbf{Ay},\mathbf{f}) + (\mathbf{f},\mathbf{f}) =$$
$$= \left(\mathbf{A}^T\mathbf{Ay},\mathbf{y}\right) - 2\left(\mathbf{A}^T\mathbf{f},\mathbf{y}\right) + (\mathbf{f},\mathbf{f}),$$

применяя необходимое (а в данном случае и достаточное) условие минимума и правила дифференцирования квадратичной и линейной форм, имеем

$$\Phi(\mathbf{y}) = \min \quad \Leftrightarrow \quad 2\mathbf{A}^T\mathbf{Ay} - 2\mathbf{A}^T\mathbf{f} = 0,$$

т.е. приходим к системе (1.24).

Найдем псевдорешение системы В. Имеем:

$$\mathbf{A} = \mathbf{A}^T = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{A}^T\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix},$$

$$\mathbf{A}^T\mathbf{f} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 0.5 \\ 1.5 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix},$$

и, значит, система (1.24) в данном случае есть

$$\begin{cases} 2x_1 + 2x_2 = 2, \\ 2x_1 + 2x_2 = 2. \end{cases} \tag{1.25}$$

Последняя фактически совпадает с системой А, имеющей, как отмечалось выше, неединственное решение. Как получить ее единственное «решение», нам уже известно.

Итак, введение понятия псевдорешения позволяет избавиться от несуществования, а понятие нормального решения — от неединственности. Объединяя их, приходим к тому, что если говорить о нахождении ***нормального псевдорешения*** (относительно задаваемого пробного решения), то первое и второе требования корректности могут быть сняты (по крайней мере, сейчас мы можем утверждать это для алгебраических систем; для более общего случая здесь заложены лишь идейные основы).

Резюмируя результат рассмотрения примера, можно сказать, что системы А и В имеют одно и то же нормальное относи-

тельно вектора $\mathbf{0} = (0;\ 0)^T$ псевдорешение $\mathbf{x}_0^* = (0.5;\ 0.5)^T$. Это решение, являясь точным решением системы Б, также служит и ее нормальным псевдорешением*).

Теперь посмотрим, можно ли полученное нормальное псевдорешение считать устойчивым.

С этой целью сначала внесем возмущение ε в один из коэффициентов системы А. Именно, наряду с системой А, имеющей нормальное псевдорешение $\mathbf{x}_0^* = (0.5;\ 0.5)^T$, будем рассматривать систему

$$\begin{cases} x_1 + x_2 = 1, \\ (1+\varepsilon)x_1 + x_2 = 1. \end{cases}$$

При любом $\varepsilon \neq 0$ определитель матрицы $\mathbf{A}_\varepsilon = \begin{pmatrix} 1 & 1 \\ 1+\varepsilon & 1 \end{pmatrix}$ этой системы равен $-\varepsilon$ $(\neq 0)$, и, следовательно, существует обратная к $\mathbf{A}_\varepsilon$ матрица $\mathbf{A}_\varepsilon^{-1} = \frac{1}{\varepsilon}\begin{pmatrix} 1 & -1 \\ -(1+\varepsilon) & 1 \end{pmatrix}$, через которую можно получить решение $\mathbf{x}_\varepsilon^*$ возмущенной системы (1.26):

$$\mathbf{x}_\varepsilon^* = \mathbf{A}_\varepsilon^{-1}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = -\frac{1}{\varepsilon}\begin{pmatrix} 0 \\ -\varepsilon \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Найденное решение $\mathbf{x}_\varepsilon^*$ оказалось независящим от возмущения ε и, вследствие этого, неприближаемым к нормальному псевдорешению $\mathbf{x}_0^*$ невозмущенной системы А.

Если такое же возмущение ε внести в систему В, т. е. наряду с ней исследовать систему

$$\begin{cases} x_1 + x_2 = 0.5, \\ (1+\varepsilon)x_1 + x_2 = 1.5, \end{cases}$$

то здесь через ту же обратную матрицу $\mathbf{A}_\varepsilon^{-1}$ получаем решение

$$\mathbf{x}_\varepsilon^* = -\frac{1}{\varepsilon}\begin{pmatrix} 1 & -1 \\ -(1+\varepsilon) & 1 \end{pmatrix}\begin{pmatrix} 0.5 \\ 1.5 \end{pmatrix} = -\frac{1}{\varepsilon}\begin{pmatrix} -1 \\ 1-0.5\varepsilon \end{pmatrix} = \begin{pmatrix} 1/\varepsilon \\ 0.5-1/\varepsilon \end{pmatrix},$$

которое при $\varepsilon \to 0$ устремляется в бесконечность, а не к нормальному псевдорешению $\mathbf{x}_0^*$ системы В.

*) К построению нормальных псевдорешений может быть применена другая техника, основанная на использовании так называемых псевдообратных матриц (Мура–Пенроуза) [43, 93].

И в том, и в другом случаях налицо *неустойчивость нормального псевдорешения по отношению к возмущению матрицы системы* (более обобщенно, к возмущению оператора).

Проследим еще за реакцией системы В на возмущение ее правой части. При любом $\varepsilon \neq 1$ система

$$\begin{cases} x_1 + x_2 = 0.5+\varepsilon, \\ x_1 + x_2 = 1.5, \end{cases} \qquad (1.27)$$

противоречива, как и исходная для нее система В. Поэтому воспользуемся ее преобразованием к виду (1.24), т. е. получим аналог системы (1.25). В данном случае имеем

$$\mathbf{A} = \mathbf{A}^T = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},\ \mathbf{f}_\varepsilon = \begin{pmatrix} 0.5+\varepsilon \\ 1.5 \end{pmatrix},\ \mathbf{A}^T\mathbf{A} = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix},\ \mathbf{A}^T\mathbf{f}_\varepsilon = \begin{pmatrix} 2+\varepsilon \\ 2+\varepsilon \end{pmatrix},$$

и, следовательно, нахождение псевдорешения системы (1.27) сводится к решению системы

$$\begin{cases} 2x_1 + 2x_2 = 2+\varepsilon, \\ 2x_1 + 2x_2 = 2+\varepsilon, \end{cases} \Leftrightarrow \begin{cases} x_1 + x_2 = 1+0.5\varepsilon, \\ x_1 + x_2 = 1+0.5\varepsilon. \end{cases}$$

Решением последней служит множество векторов вида $(x_1; 1+0.5\varepsilon - x_1)^T$, где x_1 — свободная переменная. Теперь ставим экстремальную задачу

$$x_1^2 + (1+0.5\varepsilon - x_1)^2 \to \min,$$

и из условия $2x_1 - 2(1+0.5\varepsilon - x_1) = 0$ находим значение $x_1 = 0.5+0.25\varepsilon$ первой компоненты нормального псевдорешения; вторая его компонента есть $1+0.5\varepsilon - x_1 = 0.5+0.25\varepsilon$. Так как

$$\mathbf{x}_\varepsilon^* = \begin{pmatrix} 0.5+0.25\varepsilon \\ 0.5+0.25\varepsilon \end{pmatrix} \xrightarrow[\varepsilon\to 0]{} \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} = \mathbf{x}_0^*,$$

то, по крайней мере, в этом частном случае можно говорить об устойчивости нормального псевдорешения по отношению к возмущению правой части.

Осознание того факта, что несуществование и неединственность решения уравнения (1.23) можно обойти путем обобщения (расширения) понятия решения и сужения множества, на котором ищется это обобщенное решение, привело к новой формулировке корректности.

Определение 1.4. *Задача о нахождении решения у уравнения* (1.23) *с непрерывным оператором A называется*

условно корректной *или* ***корректной по Тихонову****), *если:*

1) *априори известно, что решение* y *существует и принадлежит заданному множеству* $M \subseteq Y$ (***множеству корректности, классу корректности***);

2) *решение* y *единственно в* M;

3) *бесконечно малым вариациям* f, *не выводящим за пределы* M, *соответствуют бесконечно малые вариации* y.

Выделение множеств корректности, как правило, эффективно производится лишь в случаях привязки задачи (1.23) к конкретным приложениям, хотя, естественно, имеется ряд общих утверждений, обосновывающих такое выделение. Во многом, благодаря трудам академика А.Н. Тихонова разработана общая стратегия построения устойчивых методов решения некорректных (неустойчивых) задач в операторной форме. В ее основе лежит понятие регуляризирующего оператора или, что то же, регуляризирующего алгоритма. Дадим одно из возможных определений [41].

Определение 1.5. *Пусть* $\overline{f}$ — *фиксированное точное значение правой части уравнения* (1.23), *а* $\widetilde{f}_\delta$ — *его приближенное значение такое, что* $\rho_F\left(\overline{f}, \widetilde{f}_\delta\right) \le \delta$.

Оператор $R(f, \alpha)$ *называется* ***регуляризирующим оператором*** (***регуляризирующим алгоритмом, регуляризатором***) *для уравнения* (1.23) *в окрестности* $\overline{f}$, *если:*

1) *оператор* $R(f, \alpha)$ *определен для любых* $\widetilde{f}_\delta \in F$, $\delta \in [0, \delta_0]$ *и* $\alpha \in [0, \alpha_0]$, *где* δ_0 *и* α_0 — *некоторые предельные значения параметров* δ *и* α;

2) *существует такая зависимость* $\alpha = \alpha(\delta)$, *что* $\forall \varepsilon > 0 \quad \exists \delta(\varepsilon) > 0 : \left[\rho_F\left(\widetilde{f}_\delta, \overline{f}\right) \le \delta \le \delta(\varepsilon) \Rightarrow \rho_Y(\widetilde{y}_\alpha, \overline{y}) \le \varepsilon\right]$, *где* $\widetilde{y}_\alpha := R\left(\widetilde{f}_\delta, \alpha(\delta)\right)$ — ***регуляризованное решение уравнения*** (1.23), *а элемент* $\overline{y}$ *таков, что* $A\overline{y} = \overline{f}$; *при этом должно быть*

$$\delta \to 0 \;\Rightarrow\; \alpha \to 0, \quad \varepsilon \to 0$$

*) Ти́хонов Андрей Николаевич (1906–1993) — российский математик и геофизик. Наряду со многими другими своими научными достижениями положил начало бурному развитию теории и практики методов решения некорректных задач (1943 г.). Термин «корректность по Тихонову» принадлежит другому крупному специалисту по некорректным задачам — академику РАН Лаврентьеву Михаилу Михайловичу (род. 1932г.). Вообще, следует отметить ведущие позиции российских ученых в данной проблематике.

(*т.е. при* $\delta \to 0$ *регуляризованное решение* $\widetilde{y}_\alpha$ *должно стремиться к точному решению* $\overline{y}$).

Заметим, что, во-первых, ниоткуда не следует единственность регуляризатора для данной задачи, а во-вторых, требование $\widetilde{y}_\alpha \to \overline{y}$ при $\delta \to 0$ имеет только теоретическое значение, поскольку уровень δ погрешностей исходных данных не может быть уменьшен по желанию вычислителя.

Всякая задача вида (1.23), для которой существует регуляризирующий оператор, называется ***регуляризуемой*** (или ***регуляризируемой***), а всякий метод, порождающий такой оператор, называется ***методом регуляризации***. Фигурирующая в определении 1.5 скалярная величина $\alpha = \alpha(\delta)$ называется ***параметром регуляризации***.

Одним из наиболее универсальных способов построения регуляризирующих алгоритмов является ***метод*** α***–регуляризации Тихонова***. Суть его в следующем.

Пусть в уравнении (1.23) A — линейный вполне непрерывный оператор (т.е. любое ограниченное множество из Y переводится им в компактное множество из F), Y и F — гильбертовы пространства, и пусть вместо «точного» уравнения (1.23) известно «приближенное» уравнение

$$\widetilde{A}y = \widetilde{f}, \tag{1.28}$$

где $y \in Y$, $\widetilde{f} \in F$, а близость (1.23) и (1.28) характеризуется неравенствами

$$\left\|\widetilde{A} - A\right\| \le \xi, \quad \left\|\widetilde{f} - f\right\| \le \delta. \tag{1.29}$$

При поиске решения уравнения (1.23) на базе уравнения (1.28) вместо минимизации невязки $\left\|\widetilde{A}y - \widetilde{f}\right\|$, которая, как отмечалось в § 1.5. на неустойчивых задачах может вести себя нерегулярно, предлагается минимизировать так называемый ***сглаживающий функционал*** (***функционал Тихонова***)

$$\Phi_\alpha\left[y, \widetilde{A}, \widetilde{f}\right] := \left\|\widetilde{A}y - \widetilde{f}\right\|^2 + \alpha\,\Omega[y].$$

Здесь $\alpha > 0$ — параметр регуляризации, а $\Omega[y]$ — ***стабилизирующий функционал*** (***стабилизатор***), главное требование к нему — неотрицательность; чаще всего, берут $\Omega[y] := \|y\|^2$. Доказано, что функционал Тихонова, в частности,

$$\Phi_\alpha\left[y, \widetilde{A}, \widetilde{f}\right] := \left\|\widetilde{A}y - \widetilde{f}\right\|^2 + \alpha\|y\|^2 \tag{1.30}$$

всегда имеет и притом единственный элемент y_α такой, что

$$\Phi_\alpha\left[y_\alpha, \widetilde{A}, \widetilde{f}\right] = \inf_{y \in Y} \Phi_\alpha\left[y, \widetilde{A}, \widetilde{f}\right].$$

При всяком фиксированном $\alpha > 0$ последнюю задачу можно решать как подходящими численными методами минимизации, так и с помощью решения ***уравнения Тихонова***

$$\alpha y_\alpha + \widetilde{A}^* \widetilde{A} y_\alpha = \widetilde{A}^* \widetilde{f}, \qquad (1.31)$$

которое представляет собой запись необходимого условия экстремума $\Phi_\alpha[y, \widetilde{A}, \widetilde{f}]$ ($\widetilde{A}^*$ — оператор, сопряженный оператору $\widetilde{A}$).

Так как реально $\delta \not\to 0$, $\xi \not\to 0$, то параметр регуляризации α должен подбираться так, чтобы принимаемое за решение задачи (1.23) регуляризованное решение y_α было, по возможности, наилучшим в смысле удачного согласования левой и правой частей уравнения (1.28) (что определяется малостью невязки $\|\widetilde{A}y_\alpha - \widetilde{f}\|^2$) и устойчивости его вычисления (что связано с величиной стабилизатора $\|y_\alpha\|^2$). При этом, чем меньше δ и ξ, тем меньшим должно быть значение α, т.е. бóльшим относительный вес первого слагаемого в функционале Тихонова. Таким образом, в методе α-регуляризации Тихонова (впрочем, как и в других методах регуляризации) главной проблемой является проблема выбора параметра регуляризации, призванного сбалансировать возможность получения регуляризованного решения y_α, как можно более близкого (в условиях оговоренной неопределенности) к решению именно «нужной» задачи, с тем, чтобы оно могло быть надежно вычислено. Имеется несколько способов выбора параметра регуляризации [34, 41, 128 и др.].

Для интерпретации процедуры α-регуляризации вновь обратимся к рассматривавшимся в начале параграфа системам линейных алгебраических уравнений.

Можно считать, например, что система А служит приближенной для системы Б. Тогда в обозначениях (1.23), (1.28) имеем:

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1.00001 \end{pmatrix}, \quad \widetilde{\mathbf{A}} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} 1 \\ 1.000005 \end{pmatrix}, \quad \widetilde{\mathbf{f}} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

При этом сама система Б, вообще говоря, предполагается неизвестной; известно лишь, насколько она близка к системе А. Положив $\mathbf{y} = (x_1; x_2)^T$, находим $\|\mathbf{y}\|_2^2 = x_1^2 + x_2^2$,

$$\widetilde{\mathbf{A}}\mathbf{y} - \widetilde{\mathbf{f}} = \begin{pmatrix} x_1 + x_2 - 1 \\ x_1 + x_2 - 1 \end{pmatrix}, \quad \|\widetilde{\mathbf{A}}y - \widetilde{\mathbf{f}}\|_2^2 = 2(x_1 + x_2 - 1)^2.$$

Следовательно, функционал Тихонова (1.30) в данном случае есть

$$\Phi_\alpha[y, \widetilde{A}, \widetilde{f}] = 2(x_1 + x_2 - 1)^2 + \alpha(x_1^2 + x_2^2),$$

а α-регуляризованное по Тихонову решение —

$$\mathbf{y}_\alpha \equiv \begin{pmatrix} x_1(\alpha) \\ x_2(\alpha) \end{pmatrix} = \underset{(x_1; x_2)^T \in \mathbf{R}_2}{\arg\min} \left[2(x_1 + x_2 - 1)^2 + \alpha(x_1^2 + x_2^2)\right].$$

Так как здесь $\mathbf{A}^* = \mathbf{A}^T = \mathbf{A}$, то уравнение Тихонова (1.31) принимает вид

$$\alpha\begin{pmatrix} x_1(\alpha) \\ x_2(\alpha) \end{pmatrix} + \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}\begin{pmatrix} x_1(\alpha) \\ x_2(\alpha) \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

или после упрощения

$$\begin{pmatrix} 2+\alpha & 2 \\ 2 & 2+\alpha \end{pmatrix}\begin{pmatrix} x_1(\alpha) \\ x_2(\alpha) \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}.$$

Отсюда легко получаем параметризованное решение

$$\mathbf{y}_\alpha = \begin{pmatrix} x_1(\alpha) \\ x_2(\alpha) \end{pmatrix} = \frac{2}{4+\alpha}\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad (1.32)$$

при $\alpha \to 0$ стремящемся к нормальному псевдорешению $(0.5;\ 0.5)^T$. Однако устремлять α к нулю как раз и не следует, а нужно каким-то образом зафиксировать его так, чтобы оно отражало связь систем А и Б, характеризуемую степенью близости $\widetilde{\mathbf{A}}$ к $\mathbf{A}$, $\widetilde{\mathbf{f}}$ к $\mathbf{f}$. Как уже отмечалось, нахождение оптимальных значений параметра α в общем случае представляет собой сложную задачу. Но для линейных алгебраических систем имеются простые прикидки для значений параметра α и точности приближения нормального псевдорешения $\mathbf{y}_0$ регуляризованным решением $\mathbf{y}_\alpha$ при таких α. Например, считается [33], что если в оценках (1.29) близости систем (1.23) и (1.28) $\xi = \delta = \varepsilon$, то регуляризированное решение $\mathbf{y}_\alpha$ системы (1.28) приближает нормальное псевдорешение системы (1.23) с точностью $O\left(\varepsilon^{\frac{2}{3}}\right)$ при $\alpha = \varepsilon^{\frac{2}{3}}$, если эта система разрешима, и с точностью $O\left(\varepsilon^{\frac{1}{2}}\right)$ при $\alpha = \varepsilon^{\frac{1}{2}}$, если она неразрешима. Если в рассматриваемом примере несколько

завысить δ и посчитать, что $\xi = \delta = 0.00001$, то, согласно приведенной рекомендации, можно взять $\alpha_0 = 10^{-\frac{10}{3}} \approx 0.000464$ и, подставив это значение в формулу (1.32), получить регуляризованное решение $\mathbf{y}_{\alpha_0} \approx \begin{pmatrix} 0.499942 \\ 0.499942 \end{pmatrix}$.

В заключение параграфа отметим, что информацию о методах решения некорректных задач (в настоящее время все чаще говорят — неустойчивых) можно найти и в научно-популярной [33, 93], и в учебной [78, 148, 174, 181, 194], и, естественно, в специальной литературе [11, 28, 34, 41, 128, 175 и др.]. С некорректными задачами в этой книге мы еще встретимся далее, например, в главах 13 и 18.

УПРАЖНЕНИЯ

1.1. Найдите приближенное значение выражения

$$\frac{3.7894 \cdot 0.29^2}{5.63},$$

считая, что входящие в него числа (за исключением показателя степени) — приближенные, записанные в соответствии с правилом Крылова.

Сделайте теоретическую оценку главной части абсолютной погрешности результата.

1.2. Пусть машинное слово некой ЭВМ состоит из 16 четверичных разрядов.

А) Найдите приближенно значения машинного нуля M_0, машинной бесконечности M_∞ и машинного эпсилон в системе представления с плавающей запятой, предполагая, что под мантиссу выделяется 11 разрядов.

Б) Оцените абсолютную погрешность и диапазон представимости чисел с фиксированной запятой при условии, что под целую и дробную части числа выделяется поровну разрядов.

(Считается, что для отображения знаков «+», «–», и «,» используются те же четверичные элементы и что округление производится на основе простого отбрасывания).

1.3. Обращением матрицы коэффициентов решите систему

$$\begin{cases} x_1 \quad - \quad x_3 = 2, \\ x_1 + x_2 + x_3 = 6, \\ \quad x_2 + 3x_3 = 5. \end{cases}$$

Найдя число обусловленности, оцените возможные относительное и абсолютное отклонения от полученного решения при относительной ошибке в правой части порядка 0.01 (по некоторой норме).

1.4. Докажите справедливость неравенств (1.14). Примените одно из этих неравенств к системе упр.1.3, считая, что в ее матрицу $\mathbf{A}$ внесено возмущение $\Delta\mathbf{A}$ такое, что $\|\Delta\mathbf{A}\| = 0.01$.

1.5. Пусть коэффициенты уравнения

$$x^2 - 6.9x - 0.0212 = 0$$

являются приближенными числами, записанными в соответствии с принципом Крылова (коэффициент при x^2 и правую часть считаем числами точными).

А) Найдите корни уравнения с максимальной разумной точностью, диктуемой степенью неопределенности коэффициентов.

Б) Подсчитайте числа обусловленности для найденных корней и на этой основе дайте приближенную оценку их относительных и абсолютных погрешностей.

1.6. Найдите множество решений и нормальное (относительно нулевого вектора) решение системы

$$\begin{cases} 2x_1 + 3x_2 + x_3 = 13, \\ x_1 + x_2 + x_3 = 6, \\ 3x_1 + 5x_2 + x_3 = 20. \end{cases}$$

1.7. Убедитесь в плохой обусловленности систем:

а) $\begin{cases} 2x_1 - 5x_2 = 1, \\ 2x_1 - 4.9999x_2 = 1; \end{cases}$

б) $\begin{cases} 2x_1 - 5x_2 = 1, \\ 2x_1 - 4.9999x_2 = 1.0003; \end{cases}$

в) $\begin{cases} 2x_1 - 5x_2 = 1, \\ 2x_1 - 5x_2 = 1.0003. \end{cases}$

Найдите их нормальные псевдорешения. Принимая за основу систему

$$\begin{cases} 2x_1 - 5x_2 = 1, \\ 2x_1 - 5x_2 = 1, \end{cases}$$

получите α-регуляризованные решения систем а–в.

ГЛАВА 2 РЕШЕНИЕ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ СИСТЕМ (ПРЯМЫЕ МЕТОДЫ)

Рассматриваются простые, часто применяемые численные методы решения систем линейных алгебраических уравнений с квадратными матрицами коэффициентов. Попутно решаются задачи обращения матриц и вычисления определителей. Наряду с широко известным методом Гаусса, изучаемым здесь в плане реальных вычислений, вниманию читателя предлагаются также метод вращений (более устойчивый к погрешностям арифметических действий, чем метод Гаусса), метод квадратных корней для решения систем с симметричными матрицами и метод прогонки для систем с трехдиагональными матрицами (т.е. для трехточечных разностных уравнений второго порядка).

2.0. ВВЕДЕНИЕ

Как утверждается в книге известного американского математика Валяха [29], 75% всех расчетных математических задач приходится на решение систем линейных алгебраических уравнений (в дальнейшем, СЛАУ). Это не удивительно, так как математические модели тех или иных явлений или процессов либо сразу строятся как линейные алгебраические, либо сводятся к таковым посредством дискретизации и/или линеаризации. Поэтому трудно переоценить роль, которую играет выбор эффективного (в том или ином смысле) способа решения СЛАУ. Современная вычислительная математика располагает большим арсеналом методов, а математическое обеспечение ЭВМ — многими пакетами прикладных программ, позволяющих решать различные возникающие на практике линейные системы. Чтобы ориентироваться среди методов и программ и в нужный момент сделать оптимальный выбор, нужно разбираться в основах построений методов и алгоритмов, учитывающих специфику постановок задач, знать их сильные и слабые стороны и границы применимости.

Все методы решения линейных алгебраических задач (наряду с задачей решения СЛАУ, это и вычисление определителей, и обращение матриц, и задачи на собственные значения) можно разбить на два класса: прямые и итерационные. Как явствует из заглавия, здесь будут рассмотрены только ***прямые методы***, т.е. такие методы, которые приводят к решению за конечное число

арифметических операций. Если операции реализуются точно, то и решение также будет точным (в связи с чем к классу прямых методов применяют еще название ***точные методы***). ***Итерационным методам***, т.е. методам, в которых точное решение может быть получено лишь в результате бесконечного повторения единообразных (как правило, простых) действий, будет посвящена следующая глава.

Другое ограничение будет касаться рассматриваемых систем. Условимся говорить о численном решении таких СЛАУ, у которых число уравнений совпадает с числом вещественных неизвестных, причем будем предполагать наличие единственного решения*), если существование и единственность не следует из каких-либо условий.

Такое ограничение здесь довольно естественно, так как решение и недоопределенных, и переопределенных систем, а также систем с комплексными коэффициентами и переменными, в конечном счете, сводится к решению однозначно определенных вещественных систем.

Итак, изучается вопрос о численном решении систем вида

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2, \\ \ldots\ldots\ldots\ldots\ldots & \ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n, \end{cases} \qquad (2.1)$$

или иначе, векторно-матричных уравнений

$$\mathbf{Ax} = \mathbf{b}, \qquad (2.1а)$$

где $\mathbf{b} = (b_1, b_2, \ldots, b_n)^T$ — вектор свободных членов и $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ — вектор неизвестных (он же в другой интерпретации может означать и вектор-решение) с вещественными координатами, а $\mathbf{A} = (a_{ij})_{i,j=1}^{n}$ — вещественная $n \times n$-матрица коэффициентов данной системы. Эффективность способов решения системы (2.1) во многом зависит от структуры и свойств матрицы $\mathbf{A}$: размера, обусловленности (см. §1.5), симметрично-

*) Обратим внимание на двоякое употребление в тексте термина «решение»:

1) решение — как упорядоченный набор (вектор) значений неизвестных, удовлетворяющих всем уравнениям системы;

2) решение — как процесс нахождения решения в первом смысле.

сти, заполненности (т.е. соотношения между числом ненулевых и нулевых элементов), специфики расположения ненулевых элементов в матрице и др.

Так, размерность системы (т.е. число n) является главным фактором, заставляющим вычислителей отвернуться от весьма привлекательных в теоретическом плане и приемлемых на практике при небольших n (2, 3) ***формул Крамера***

$$x_i = \frac{\det \mathbf{A}_i}{\det \mathbf{A}} \quad (i = 1, 2, \ldots, n),$$

позволяющих находить неизвестные компоненты вектора $\mathbf{x}$ в виде дробей, знаменателем которых является определитель матрицы системы, а числителем — определители матриц $\mathbf{A}_i$, полученные из $\mathbf{A}$ заменой столбца коэффициентов при вычисляемом неизвестном столбцом свободных членов. Если при реализации этих формул определители вычисляются понижением порядка на основе разложения по элементам какой-нибудь строки или столбца матрицы, то на вычисление определителя n-го порядка будет затрачиваться $n!$ операций умножения. Факториальный рост количества арифметических операций (и вообще, очень быстрый рост) с увеличением размерности задачи называют «проклятьем размерности». Что это такое, можно представить, зафиксировав, например, $n = 100$. Оценив величину $100! \approx 10^{158}$ и прикинув потенциальные возможности развития вычислительной техники, приходим к выводу о том, что в обозримом будущем системы сотого порядка в принципе не могут быть решены по формулам Крамера [12, 92]. Заметим при этом, что, во-первых, метод Крамера будет неустойчив, т.е. погрешности округлений будут катастрофически нарастать, во-вторых, размерность $n = 100$ для современных задач не так и велика: довольно часто решаются системы с сотнями и с тысячами неизвестных.

Если осуществлять вычисление обратной матрицы с помощью союзной матрицы, т. е. через алгебраические дополнения, то нахождение решения векторно-матричного уравнения (2.1а) по формуле

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

фактически равнозначно применению формул Крамера и также практически непригодно по упомянутым выше причинам для вычислительных целей.

2.1. АЛГОРИТМ РЕШЕНИЯ СЛАУ МЕТОДОМ ГАУССА С ПОСТОЛБЦОВЫМ ВЫБОРОМ ГЛАВНОГО ЭЛЕМЕНТА

Наиболее известным и популярным способом решения линейных систем вида (2.1) является метод Гаусса. Суть его проста — это последовательное исключение неизвестных. В отличие от курсов линейной алгебры, нас будут интересовать вычислительные аспекты метода Гаусса, а именно, технология получения вектора-решения $\mathbf{x}$ из исходных матрицы $\mathbf{A}$ и вектора $\mathbf{b}$, причем, по возможности, минимизирующая влияние неизбежных ошибок округлений. С этой целью, работая с уравнениями системы (2.1), выведем сначала совокупность формул, позволяющих в итоге получить искомые значения неизвестных, а затем на их основе запишем алгоритм решения поставленной задачи.

Будем поэтапно приводить систему (2.1) к треугольному виду, исключая последовательно сначала x_1 из второго, третьего, ... , n-го уравнений, затем x_2 из третьего, четвертого, ... , n-го уравнений преобразованной системы, и т.д.

На первом этапе заменим второе, третье, ... , n-е уравнения на уравнения, получающиеся сложением этих уравнений с первым, умноженным соответственно на $-\frac{a_{21}}{a_{11}}, -\frac{a_{31}}{a_{11}}, \ldots, -\frac{a_{n1}}{a_{11}}$. Результатом этого этапа преобразований будет эквивалентная (2.1) система

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1, \\ \qquad a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)}, \\ \qquad a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 + \ldots + a_{3n}^{(1)}x_n = b_3^{(1)}, \\ \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \qquad a_{n2}^{(1)}x_2 + a_{n3}^{(1)}x_3 + \ldots + a_{nn}^{(1)}x_n = b_n^{(1)}, \end{cases} \tag{2.2}$$

коэффициенты которой (с верхним индексом 1) подсчитываются по формулам

$$a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}}{a_{11}}a_{1j}, \quad b_i^{(1)} = b_i - \frac{a_{i1}}{a_{11}}b_1, \quad \text{где } i, j = 2, 3, \ldots, n.$$

При этом можно считать, что $a_{11} \neq 0$, так как по предположению система (2.1) однозначно разрешима, значит, все коэффициенты при x_1 не могут одновременно равняться нулю, и на первое место всегда можно поставить уравнение с отличным от нуля первым коэффициентом.

На втором этапе проделываем такие же операции, как и на первом, с подсистемой системы (2.2), получающейся исключением первого уравнения. Эквивалентный (2.2) (а значит, и (2.1)) результат второго этапа будет иметь вид

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1 , \\ \qquad a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)}, \\ \qquad\qquad a_{33}^{(2)}x_3 + \ldots + a_{3n}^{(2)}x_n = b_3^{(2)}, \\ \qquad\qquad \ldots\ldots\ldots\ldots\ldots \\ \qquad\qquad a_{n3}^{(2)}x_3 + \ldots + a_{nn}^{(2)}x_n = b_n^{(2)}, \end{cases}$$

где $a_{ij}^{(2)} = a_{ij}^{(1)} - \dfrac{a_{i2}^{(1)}}{a_{22}^{(1)}} a_{2j}^{(1)}$, $b_i^{(2)} = b_i^{(1)} - \dfrac{a_{i2}^{(1)}}{a_{22}^{(1)}} b_2^{(1)}$; $i, j = 3, \ldots, n$.

Продолжая этот процесс, на $(n-1)$-м этапе так называемого ***прямого хода*** метода Гаусса данную систему (2.1) приведем к треугольному виду:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 , \\ \qquad a_{22}^{(1)}x_2 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)}, \\ \qquad\qquad \ldots\ldots\ldots\ldots \\ \qquad\qquad a_{nn}^{(n-1)}x_n = b_n^{(n-1)}. \end{cases} \tag{2.3}$$

На основе предыдущих рассуждений и формул легко убедиться, что коэффициенты этой системы могут быть получены из коэффициентов данной системы последовательным пересчетом по формулам

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} a_{kj}^{(k-1)}, \quad b_i^{(k)} = b_i^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} b_k^{(k-1)}, \tag{2.4}$$

где верхний индекс k (номер этапа) должен изменяться от 1 до $n-1$, нижние индексы i и j (в любой очередности) — от k+1 до n; по определению полагаем $a_{ij}^{(0)} := a_{ij}$, $b_i^{(0)} := b_i$.

Треугольная, точнее, трапециевидная структура системы (2.3) позволяет последовательно одно за другим вычислять значения неизвестных, начиная с последнего:

$$x_n = \frac{b_n^{(n-1)}}{a_{nn}^{(n-1)}};$$

$$\ldots\ldots\ldots$$

$$x_2 = \frac{b_2^{(1)} - a_{23}^{(1)}x_3 - \ldots - a_{2n}^{(1)}x_n}{a_{22}^{(1)}};$$

$$x_1 = \frac{b_1 - a_{12}x_2 - \ldots - a_{1n}x_n}{a_{11}}.$$

Этот процесс последовательного вычисления значений неизвестных называют ***обратным ходом*** метода Гаусса. Очевидно, он определяется одной формулой

$$x_k = \frac{1}{a_{kk}^{(k-1)}} \left(b_k^{(k-1)} - \sum_{j=k+1}^{n} a_{kj}^{(k-1)} x_j \right), \tag{2.5}$$

где k полагают равным $n, n-1, \ldots, 2, 1$ и сумма по определению считается равной нулю, если нижний предел суммирования у знака Σ имеет значение больше верхнего.

Итак, решение СЛАУ вида (2.1) методом Гаусса сводится к последовательной реализации вычислений по формулам (2.4) и (2.5).

Учитывая цикличность выполняемых при этом операций, а также нецелесообразность хранения промежуточных результатов (пересчитываемых коэффициентов промежуточного этапа), запишем простой **алгоритм решения линейных систем (2.1) методом Гаусса** [138]:

1. для $k = 1, 2, \ldots, n-1$,
2. для $i = k+1, \ldots, n$:
3. $t_{ik} := a_{ik} / a_{kk}$,
4. $b_i := b_i - t_{ik} b_k$;
5. для $j = k+1, \ldots, n$:
6. $a_{ij} := a_{ij} - t_{ik} a_{kj}$.
7. $x_n := b_n / a_{nn}$;
8. для $k = n-1, \ldots, 2, 1$:
9. $x_k := \left(b_k - \sum_{j=k+1}^{n} a_{kj} x_j \right) / a_{kk}$.

Подав на его вход квадратную матрицу $(a_{ij})_{i,\,j=1}^{n}$ коэффициентов при неизвестных системы (2.1) и вектор $(b_i)_{i=1}^{n}$ свободных членов и выполнив три вложенных цикла вычислений прямого хода (строки 1–6) и один цикл вычислений обратного хода (строки 7–9), на выходе алгоритма получим вектор-решение $(x_k)_{k=1}^{n}$ (в обратном порядке), если, разумеется, ни один из знаменателей не обращается в нуль и все вычисления проводятся точно.

Так как реальные машинные вычисления производятся не с точными, а с усеченными числами, т.е. неизбежны ошибки округления, то анализируя, например, формулы (2.4), можно сделать вывод о том, что выполнение алгоритма может прекратиться или привести к неверным результатам, если знаменатели дробей на каком-то этапе окажутся равными нулю или очень маленькими числами. Чтобы уменьшить влияние ошибок округлений и исключить деление на нуль, на каждом этапе прямого хода уравнения системы (точнее, обрабатываемой подсистемы) обычно переставляют так, чтобы деление производилось на наибольший по модулю в данном столбце (обрабатываемом подстолбце) элемент. Числа, на которые производится деление в методе Гаусса, называются ***ведущими*** или ***главными элементами.*** Отсюда название рассматриваемой модификации метода, исключающей деление на нуль и уменьшающей вычислительные погрешности, — ***метод Гаусса с постолбцовым выбором главного элемента*** (или, иначе, ***с частичным упорядочиванием по столбцам***).

Частичное упорядочивание по столбцам требует внесения в алгоритм следующих изменений: между строками 1 и 2 нужно сделать вставку*)

⊗ «Найти $m \ge k$ такое, что $|a_{mk}| = \max\limits_{i \ge k}\{|a_{ik}|\}$;
если $a_{mk} = 0$, остановить работу алгоритма («однозначного решения нет»),
иначе поменять местами b_k и b_m, a_{kj} и a_{mj} при всех $j = k, \ldots, n$.»

Более разумным, наверное, является сравнение $|a_{mk}|$ не с нулем, а с некоторым малым $\varepsilon > 0$, задаваемым вычислителем в зависимости от различных априорных соображений. Счет останавливается или берется под особый контроль, если окажется $|a_{mk}| < \varepsilon$. Заметим, что соответствующая частичному упорядочиванию вставка ⊗ в алгоритм позволяет фактически в процессе его выполнения проводить алгоритмическое исследование системы (2.1) на однозначную разрешимость.

Устойчивость алгоритма к погрешностям исходных данных и результатов промежуточных вычислений можно еще усилить, если выполнять деление на каждом этапе на элемент, наибольший по модулю в о в с е й м а т р и ц е преобразуемой на данном этапе подсистемы. Такая модификация метода Гаусса, называемая ***методом главных элементов***, применяется довольно редко, поскольку сильно усложняет алгоритм. Усложнение связано как с необходимостью осуществления двумерного поиска главных элементов, так и с необходимостью запоминать номера столбцов, откуда берутся эти элементы (перестановка столбцов означает как бы переобозначение неизвестных, в связи с чем требуется обратная замена).

*) Чтобы частичное упорядочивание было более эффективным, перед этим целесообразно произвести ***масштабирование*** (уравновешивание) системы, например, разделить все числа каждой строки на наибольшее число строки [138].

2.2. ПРИМЕНЕНИЕ МЕТОДА ГАУССА К ВЫЧИСЛЕНИЮ ОПРЕДЕЛИТЕЛЕЙ И К ОБРАЩЕНИЮ МАТРИЦ

Как было сказано в § 2.0, решения линейных алгебраических систем можно получать с помощью определителей или обратных матриц. Однако нетрудно увидеть, что более эффективно поступать наоборот: вычислять определители и обращать матрицы в рамках метода Гаусса решения линейных систем.

Действительно, выполняемые в методе Гаусса преобразования прямого хода, приведшие матрицу $\mathbf{A}$ системы к треугольному виду (см. (2.3)) таковы, что они не изменяют определителя матрицы $\mathbf{A}$. Учитывая, что определитель треугольной матрицы равен произведению диагональных элементов, имеем

$$\det \mathbf{A} = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ 0 & a_{22}^{(1)} & \ldots & a_{2n}^{(1)} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & a_{nn}^{(n-1)} \end{vmatrix} = a_{11} \cdot a_{22}^{(1)} \cdot \ldots \cdot a_{nn}^{(n-1)} .$$

Таким образом, *определитель матрицы равен произведению всех ведущих элементов при ее преобразовании методом Гаусса.*

При желании получить $\det \mathbf{A}$ дополнительно к решению СЛАУ $\mathbf{Ax} = \mathbf{b}$ алгоритм предыдущего пункта должен быть пополнен всего лишь одной следующей строкой:

10. $\det \mathbf{A} = \prod_{k=1}^{n} a_{kk}$.

Если метод Гаусса используется только для вычисления определителя, из алгоритма его реализации следует изъять строки 4 и 7–9.

Так как перестановка строк матрицы меняет знак определителя, то при постолбцовом выборе главного элемента, т. е. при включении в алгоритм вставки ⊗, нужно в результате учесть число p произведенных перестановок, точнее, четность этого числа. Это означает, что при вычислении $\det \mathbf{A}$ алгоритмом Гаусса с частичным упорядочиванием вместо строки 10 должна быть включена строка

10'. $\det \mathbf{A} = (-1)^p \prod_{k=1}^{n} a_{kk}$.

Пример 2.1. Дана матрица $\mathbf{A} = \begin{pmatrix} 2 & -1 & 1 \\ 4 & 3 & 1 \\ 6 & -13 & 6 \end{pmatrix}$. Методом Гаусса найти $\det \mathbf{A}$.

Для данного случая алгоритм предыдущего параграфа может быть конкретизирован так:

1. для $k = 1, 2$,
2. для $i = k+1, 3$:
3. $t_{ik} := \dfrac{a_{ik}}{a_{kk}}$
4. для $j = k+1, 3$:
5. $a_{ij} := a_{ij} - t_{ik} a_{kj}$;
6. $\det \mathbf{A} := a_1 \cdot a_2 \cdot a_3$.

Воспользуемся им. Фиксируем $k = 1$ и, соответственно, $i = 2$. Тогда $t_{21} := \dfrac{a_{21}}{a_{11}} = \dfrac{4}{2} = 2$, и при $j = 2, 3$ имеем

$$a_{22} := a_{22} - t_{21} a_{12} = 3 - 2(-1) = 5,$$
$$a_{23} := a_{23} - t_{21} a_{13} = 1 - 2 \cdot 1 = -1.$$

Переключая i на значение 3, далее подсчитываем $t_{31} := \dfrac{a_{31}}{a_{11}} = \dfrac{6}{2} = 3$ и при $j = 2, 3$ получаем

$$a_{32} := a_{32} - t_{31} a_{12} = -13 - 3(-1) = -10,$$
$$a_{33} := a_{33} - t_{31} a_{13} = 6 - 3 \cdot 1 = 3.$$

Теперь полагаем $k = 2$ (т.е. переходим ко второму этапу преобразований). При этом значении k индексы i и j могут принять только одно значение 3. Следовательно, достаточно подсчитать

$$t_{32} := \frac{a_{32}}{a_{22}} = \frac{-10}{5} = -2 \quad \text{и} \quad a_{33} := a_{33} - t_{32} a_{23} = 3 - (-2)(-1) = 1$$

(где вместо элементов a_{22}, a_{23}, a_{32}, a_{33} исходной матрицы подставляются новые, подсчитанные на первом этапе, значения), чтобы получить искомое значение определителя

$$\det \mathbf{A} = 2 \cdot 5 \cdot 1 = 10.$$

Для получения матрицы $\mathbf{A}^{-1}$, обратной к матрице $\mathbf{A} = (a_{ij})_{i,j=1}^{n}$, будем исходить из того, что она является решением матричного уравнения

$$\mathbf{AX} = \mathbf{E}, \tag{2.6}$$

где $\mathbf{E}$ — единичная матрица.

Представляя искомую матрицу $\mathbf{X} = (x_{ij})_{i,j=1}^{n}$ как набор (вектор-строку) векторов-столбцов

$$\mathbf{x}_1 = \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix}, \quad \mathbf{x}_2 = \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix}, \quad \ldots, \quad \mathbf{x}_n = \begin{pmatrix} x_{1n} \\ x_{2n} \\ \vdots \\ x_{nn} \end{pmatrix},$$

а единичную матрицу $\mathbf{E}$ как набор единичных векторов

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots, \quad \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix},$$

матричное уравнение (2.6) в соответствии с правилами умножения матриц подменим эквивалентной системой не связанных между собой векторно-матричных уравнений

$$\mathbf{Ax}_1 = \mathbf{e}_1; \quad \mathbf{Ax}_2 = \mathbf{e}_2; \quad \ldots\ ; \quad \mathbf{Ax}_n = \mathbf{e}_n. \tag{2.7}$$

Каждое из последних уравнений имеет вид (2.1а) и может быть решено методом Гаусса. При этом специфичным является то об-

стоятельство, что все СЛАУ (2.7) имеют одну и ту же матрицу коэффициентов, а это означает, что наиболее трудоемкая часть метода Гаусса — приведение матрицы системы к треугольному виду — общая для всех систем (2.7). Так что, если требуется приспособить рассмотренный выше алгоритм решения СЛАУ методом Гаусса к обращению матриц, целесообразно не просто применить его последовательно n раз к системам (2.7), а слегка подкорректировать: «размножить» строки 4 и 9 так, чтобы в роли вектора $\mathbf{b}$ оказались все единичные векторы $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$. Тогда в результате завершения работы алгоритма будут получаться столбец за столбцом (столбцы «перевернуты»!) элементы обратной матрицы $\mathbf{X} = \mathbf{A}^{-1}$. При этом введение в алгоритм частичного упорядочивания, т.е. постолбцовый выбор главного элемента, не требует запоминаний и обратных замен.

2.3. LU-РАЗЛОЖЕНИЕ МАТРИЦ

Пусть $\mathbf{A} = (a_{ij})_{i,\,j=1}^{n}$ — данная $n \times n$-матрица, а $\mathbf{L} = (l_{ij})_{i,\,j=1}^{n}$ и $\mathbf{U} = (u_{ij})_{i,\,j=1}^{n}$ — соответственно нижняя (левая) и верхняя (правая) треугольные матрицы*). Справедливо следующее утверждение.

Теорема 2.1 [42, 158, 183]. *Если все главные миноры квадратной матрицы* $\mathbf{A}$ *отличны от нуля, то существуют такие нижняя* $\mathbf{L}$ *и верхняя* $\mathbf{U}$ *треугольные матрицы, что* $\mathbf{A} = \mathbf{LU}$. *Если элементы диагонали одной из матриц* $\mathbf{L}$ *или* $\mathbf{U}$ *фиксированы (ненулевые), то такое разложение единственно.*

Вместо полного доказательства этой теоремы (см., например, [42]) получим формулы для фактического разложения матриц в случае фиксирования диагонали нижней треугольной матрицы $\mathbf{L}$.

*) Общепринятые обозначения **L** и **U** связаны с английскими словами *lower* (нижний) и *upper* (верхний). Существует другой стандарт обозначения: **L** и **R**, определяемый словами *left* (левый) и *right* (правый).

Будем находить l_{ij} (при $i > j$) и u_{ij} (при $i \le j$) такие, чтобы

$$\begin{pmatrix} 1 & 0 & \ldots & 0 \\ l_{21} & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ l_{n1} & l_{n2} & \ldots & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & \ldots & u_{1n} \\ 0 & u_{22} & \ldots & u_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & u_{nn} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}.$$

Выполнив перемножение матриц, на основе поэлементного приравнивания левых и правых частей приходим к $n \times n$-матрице уравнений

$$\begin{array}{llll} u_{11} = a_{11}, & u_{12} = a_{12}, & \ldots, & u_{1n} = a_{1n}, \\ l_{21}u_{11} = a_{21}, & l_{21}u_{12} + u_{22} = a_{22}, & \ldots, & l_{21}u_{1n} + u_{2n} = a_{2n}, \\ \ldots & \ldots & \ldots & \ldots \\ l_{n1}u_{11} = a_{n1}, & l_{n1}u_{12} + l_{n2}u_{22} = a_{n2}, & \ldots, & l_{n1}u_{1n} + \ldots + u_{nn} = a_{nn}, \end{array}$$

относительно $n \times n$-матрицы неизвестных

$$\begin{array}{l} u_{11},\ u_{12},\ \ldots,\ u_{1n}, \\ l_{21},\ u_{22},\ \ldots,\ u_{2n}, \\ \ldots\ \ldots\ \ldots\ \ldots \\ l_{21},\ l_{2n},\ \ldots,\ u_{nn}. \end{array} \tag{2.8}$$

Специфика этой системы позволяет находить неизвестные (2.8) одно за другим в следующем порядке.

Из первой строки уравнений имеем

$$u_{1j} = a_{1j} \qquad (j = 1, \ldots, n);$$

из оставшейся части первого столбца уравнений

$$l_{i1} = \frac{a_{i1}}{u_{11}} \quad (i = 2, \ldots, n);$$

из оставшейся части второй строки

$$u_{2j} = a_{2j} - l_{21}u_{1j} \quad (j = 2, \ldots, n);$$

из оставшейся части второго столбца

$$l_{i2} = \frac{a_{i2} - l_{i1}u_{12}}{u_{22}} \quad (i = 3, \ldots, n)$$

и т.д. Последним находится элемент

$$u_{nn} = a_{nn} - \sum_{k=1}^{n-1} l_{nk}u_{kn}.$$

Легко видеть, что все отличные от 0 и 1 элементы матриц **L** и **U** могут быть однозначно вычислены с помощью всего двух формул:

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj} \quad (\text{где } i \le j), \tag{2.9}$$

$$l_{ij} = \frac{1}{u_{jj}}\left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} u_{kj} \right) \quad (\text{где } i > j). \tag{2.10}$$

При практическом выполнении разложения*) матрицы **A** нужно иметь в виду следующие два обстоятельства.

Во-первых, организация вычислений по формулам (2.9)–(2.10) должна предусматривать переключение счета с одной формулы на другую в соответствии с показанным выше процессом получения неизвестных, приведшим к этим формулам. Это удобно делать, ориентируясь на матрицу неизвестных (2.8) (ее, кстати, можно интерпретировать как n^2-мерный массив для компактного хранения LU-разложения в памяти компьютера), а именно, первая строка (2.8) вычисляется по формуле (2.9) при $i=1$, $j=1,2,\dots,n$; первый столбец (2.8) (без первого элемента) — по формуле (2.10) при $j=1$, $i=2,\dots,n$, и т.д.

Во-вторых, препятствием для осуществимости описанного процесса LU-разложения матрицы **A** может оказаться равенство нулю диагональных элементов матрицы **U**, поскольку на них выполняется деление в формуле (2.10). Отсюда требование теоремы, накладываемое на главные миноры (напомним, что главными минорами матрицы $\mathbf{A} = (a_{ij})_{i,j=1}^{n}$ называются определители подматриц $\mathbf{A}_k := (a_{ij})_{i,j=1}^{k}$, где $k=1,2,\dots,n-1$). Заметим, что $u_{11} = a_{11}$, т.е. первый диагональный элемент матрицы **U** совпадает с первым главным минором **A** и по условию должен быть отличным от нуля. Второй диагональный элемент матрицы **U**

$$u_{22} = a_{22} - l_{21}u_{12} = a_{22} - \frac{a_{21}}{a_{11}}a_{12} = \frac{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}}{a_{11}}$$

не равен нулю, если отличен от нуля второй главный минор, и т.д. Ясно, что вместо проверки на равенство нулю главных миноров данной матрицы удобнее делать такую проверку для элементов u_{jj} в процессе их вычисления, причем, чтобы уменьшить влияние погрешностей округлений, лучше сравнивать модули u_{jj} с малой положительной константой (допуском). Для определенных классов матриц требования теоремы о разложении заведомо выполняются. Это относится, например, к ***матрицам с диагональным преобладанием***, т.е. к таким, для которых

$$|a_{ii}| > \sum_{\substack{j=1 \\ (j \ne i)}}^{n} |a_{ij}| \qquad \forall i \in \{1,2,\dots,n\}.$$

*) Применяют также термины ***факторизация***, ***декомпозиция***.

2.4. РЕШЕНИЕ ЛИНЕЙНЫХ СИСТЕМ И ОБРАЩЕНИЕ МАТРИЦ С ПОМОЩЬЮ LU-РАЗЛОЖЕНИЯ

Если матрица **A** исходной системы (2.1) разложена в произведение треугольных **L** и **U**, то, значит, вместо (2.1a) можно записать эквивалентное (2.1) уравнение

$$\mathbf{LUx} = \mathbf{b}.$$

Введя вектор вспомогательных переменных $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$, последнее можно переписать в виде системы

$$\begin{cases} \mathbf{Ly} = \mathbf{b}, \\ \mathbf{Ux} = \mathbf{y}. \end{cases}$$

Таким образом, решение данной системы с квадратной матрицей коэффициентов свелось к последовательному решению двух систем с треугольными матрицами коэффициентов.

Получим сначала формулы для вычисления элементов y_i вспомогательного вектора **y**. Для этого запишем уравнение $\mathbf{Ly} = \mathbf{b}$ в развернутом виде:

$$\begin{cases} y_1 & & & = b_1, \\ l_{21}y_1 & + \; y_2 & & = b_2, \\ \dots & \dots & \dots & \dots \\ l_{n1}y_1 & + \; l_{n2}y_2 & + \dots + l_{n,n-1}y_{n-1} + y_n & = b_n. \end{cases}$$

Очевидно, все y_i могут быть последовательно найдены при $i = 1, 2, \dots, n$ по формуле

$$y_i = b_i - \sum_{k=1}^{i-1} l_{ik} y_k. \tag{2.11}$$

Развернем теперь векторно-матричное уравнение $\mathbf{Ux} = \mathbf{y}$:

$$\begin{cases} u_{11}x_1 + u_{12}x_2 + \ \dots \ + u_{1n}x_n = y_1, \\ \qquad\quad u_{22}x_2 + \ \dots \ + u_{2n}x_n = y_2, \\ \quad \dots\dots\dots\dots\dots\dots \\ \qquad\qquad\qquad\qquad u_{nn}x_n = y_n. \end{cases} \tag{2.12}$$

Отсюда значения неизвестных x_i находятся в обратном порядке, т.е. при $i = n, n-1, \dots, 2, 1$, по формуле

$$x_i = \frac{1}{u_{ii}}\left(y_i - \sum_{k=i+1}^{n} u_{ik}x_k \right). \tag{2.13}$$

Итак, решение СЛАУ посредством LU-факторизации сводится к организации вычислений по четырем формулам: совокупности формул (2.9), (2.10) для получения матрицы $\mathbf{L} + \mathbf{U} - \mathbf{E}$ (2.8) ненулевых и неединичных элементов матриц для $\mathbf{L}$ и $\mathbf{U}$, формулы (2.11) для получения вектора свободных членов треугольной системы (2.12) и формулы (2.13), генерирующей решение исходной системы (2.1).

Обратим внимание на тот факт, что выполнение расчетов по формулам (2.9)–(2.11) можно интерпретировать как преобразование системы (2.1) к треугольной системе (2.12). С системой (2.3) — результатом прямого хода метода Гаусса — последняя имеет не только структурное сходство, но полностью совпадает с ней (совпадение первых уравнений очевидно, коэффициенты при неизвестных и свободные члены вторых уравнений легко выражаются одинаково через исходные данные; можно и дальше проводить сравнение систем (2.3) и (2.12) таким путем, однако обычно идентичность этих систем показывают проще на основе представления прямого хода метода Гаусса как последовательности умножений матрицы $\mathbf{A}$ слева на матрицы очень простой структуры такой, что в итоге получается матрица $\mathbf{U}$, а последовательность обратных преобразований дает $\mathbf{L}$ [3, 74, 158, 183]. Так что решение линейных систем с помощью LU-разложения — это просто другая схема реализации метода Гаусса. В отличие от рассмотренной в § 2.1 так называемой схемы ***единственного деления*** эту называют ***компактной схемой Гаусса*** [43, 180] или ***схемой Холецкого****) [61].

*) Андре–Луи Холецкий (1875–1918) — французский военный геодезист.

Чаще схемой Холецкого называют описываемый в следующем параграфе основанный на той же идее способ решения с и м м е т р и ч н ы х линейных систем (метод квадратных корней).

Вычисление определителя LU-факторизованной матрицы $\mathbf{A}$ опирается на свойство определителя произведения матриц и сводится к перемножению n чисел:

$$\det \mathbf{A} = \det \mathbf{L} \cdot \det \mathbf{U} = u_{11} \cdot u_{22} \cdot \ldots \cdot u_{nn}.$$

Для обращения матрицы $\mathbf{A}$ с помощью LU-факторизации можно применить тот же прием, который рассмотрен в § 2.2, т.е. n-кратно использовать формулы (2.11) и (2.13) для получения столбцов матрицы $\mathbf{A}^{-1}$; при этом в качестве b_i в (2.11) должны фигурировать только 0 или 1: для нахождения первого столбца $\mathbf{A}^{-1}$ полагаем $b_1 = 1$, $b_2 = b_3 = \ldots = b_n = 0$, для второго — $b_2 = 1$, $b_1 = b_3 = \ldots = b_n = 0$, и т.д. Можно однако вывести и специальные формулы для выражения элементов обратной матрицы через элементы матриц $\mathbf{L}$ и $\mathbf{U}$. Продемонстрируем это.

Пусть матрицы $\mathbf{A}$ и $\mathbf{U}$ обратимы (матрица $\mathbf{L}$ обратима всегда). Тогда

$$\mathbf{A} = \mathbf{L} \cdot \mathbf{U} \quad \Leftrightarrow \quad \mathbf{A}^{-1} = \mathbf{U}^{-1} \cdot \mathbf{L}^{-1}.$$

Умножая последнее равенство поочередно на $\mathbf{U}$ слева и на $\mathbf{L}$ справа, будем иметь

$$\mathbf{UA}^{-1} = \mathbf{L}^{-1} \quad \text{и} \quad \mathbf{A}^{-1}\mathbf{L} = \mathbf{U}^{-1}. \tag{2.14}$$

Обозначим, как и ранее, искомые элементы матрицы $\mathbf{A}^{-1}$ через x_{ij}. Учитывая, что треугольные матрицы при обращении сохраняют свою структуру, перепишем равенства (2.14) в следующем виде

$$\begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & u_{nn} \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ * & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ * & * & \cdots & 1 \end{pmatrix},$$

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{pmatrix} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ l_{21} & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ l_{n1} & l_{n2} & \cdots & 1 \end{pmatrix} = \begin{pmatrix} * & * & \cdots & * \\ 0 & * & \cdots & * \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & * \end{pmatrix}.$$

Звездочкой здесь обозначены некоторые числа, указывающие на структуру матрицы (знание которых для дальнейшего не требуется). Полученные матричные равенства можно рассматривать как систему $2n^2$ уравнений с n^2 неизвестными x_{ij}

$(i, j = 1, 2, \ldots, n)$. Из этих $2n^2$ уравнений ровно n^2 имеют известные правые части (0 или 1). Выпишем соответствующую им $n \times n$-матрицу уравнений:

$$\begin{array}{llll} u_{11}x_{11}+\ldots+u_{1n}x_{n1}=1, & u_{11}x_{12}+\ldots+u_{1n}x_{n2}=0, & \ldots, & u_{11}x_{1n}+\ldots+u_{1n}x_{nn}=0, \\ x_{21}+\ldots+x_{2n}l_{n1}=0, & u_{22}x_{22}+\ldots+u_{2n}x_{n2}=1, & \ldots, & u_{22}x_{2n}+\ldots+u_{2n}x_{nn}=0, \\ \ldots\ldots\ldots & \ldots\ldots\ldots & \ldots & \ldots\ldots\ldots \\ x_{n1}+\ldots+x_{nn}l_{n1}=0, & x_{n2}+\ldots+x_{nn}l_{n2}=0, & \ldots, & u_{nn}x_{nn}=1. \end{array}$$

Короче все эти уравнения могут быть представлены следующими тремя типами связей*):

$$\sum_{k=i}^{n} u_{ik} x_{kj} = 0, \quad \text{если } i < j,$$

$$\sum_{k=i}^{n} u_{ik} x_{kj} = 1, \quad \text{если } i = j$$

и

$$x_{ij} + \sum_{k=j+1}^{n} x_{ik} l_{kj} = 0, \quad \text{если } i > j.$$

Отсюда можно выразить все элементы x_{ij} искомой обратной матрицы $\mathbf{A}^{-1}$:

$$x_{jj} = \frac{1}{u_{jj}} \left(1 - \sum_{k=j+1}^{n} u_{jk} x_{kj} \right); \qquad (2.15)$$

$$x_{ij} = -\frac{1}{u_{ii}} \sum_{k=i+1}^{n} u_{ik} x_{kj} \, (i < j); \qquad (2.16)$$

$$x_{ij} = -\sum_{k=j+1}^{n} x_{ik} l_{kj} \, (i > j). \qquad (2.17)$$

Формулы (2.15)–(2.17) позволяют эффективно обращать

---

*) Используя *символ Кронекера* $\delta_{ij} := \begin{cases} 1, & \text{если } i = j, \\ 0, & \text{если } i \neq j \end{cases}$, первые две формулы можно совместить:

$$\sum_{k=i}^{n} u_{ik} x_{kj} = \delta_{ij}.$$

Это относится и к формулам (2.15), (2.16).

LU-факторизованную матрицу, если соблюдать определенную технологию их использования. А именно, как видно из записанной выше матрицы уравнений, следует сначала из последнего с т о л б ц а уравнений найти x_{nn}, $x_{n-1,n}$, …, x_{2n}, x_{1n}, затем из оставшейся части последней с т р о к и уравнений найти $x_{n,n-1}, \ldots, x_{n2}, x_{n1}$, потом переключиться на предпоследний с т о л б е ц и т.д. Схематично последовательность вычислений элементов обратной матрицы можно изобразить пронумерованными стрелками следующим образом:

x_{11} … $x_{1,n-1}$ x_{1n}
3 1
$x_{n-1,1}$ … $x_{n-1,n-1}$ $x_{n-1,n}$
4
x_{n1} … $x_{n,n-1}$ x_{nn}
2

При этом стрелка 1 означает, что фиксируем $j = n$ и ведем счет по формулам (2.15), (2.16) при $i = n, n-1, \ldots, 1$; стрелка 2 — счет по формуле (2.17) при $i = n$ и $j = n-1, n-2, \ldots, 1$, и т.д.

Возвращаясь к началу § 2.3, заметим, что так же употребительно фиксирование единичной диагонали у правой треугольной матрицы, т.е. представление матрицы $\mathbf{A}$ в виде

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{pmatrix} \cdot \begin{pmatrix} 1 & u_{12} & \cdots & u_{1n} \\ 0 & 1 & \cdots & u_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

В этом случае l_{ij} и u_{ij} находятся по формулам

$$l_{ij} = a_{ij} - \sum_{k=1}^{j-1} l_{ik} u_{kj} \qquad (i \geq j),$$

$$u_{ij} = \frac{1}{l_{ii}} \left(a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj} \right) \qquad (i < j),$$

где индексы фиксируются так, чтобы вычислялись поочередно столбец $(l_{i1})_{i=1}^{n}$, затем строка $(u_{1j})_{j=2}^{n}$ и т.д.

Решение системы (2.1) с таким образом факторизованной матрицей коэффициентов получают по формулам

$$y_i = \frac{1}{l_{ii}}\left(b_i - \sum_{k=1}^{i-1} l_{ik} y_k\right), \quad i = 1, 2, \ldots, n,$$

$$x_i = y_i - \sum_{k=i+1}^{n} u_{ik} x_k, \quad i = n, n-1, \ldots, 1.$$

Детерминант матрицы $\mathbf{A}$ равен произведению $l_{11} l_{22} \ldots l_{nn}$, а для подсчета элементов обратной матрицы используют совокупность формул

$$x_{ii} = \frac{1}{l_{ii}}\left(1 - \sum_{k=i+1}^{n} x_{ik} l_{ki}\right),$$

$$x_{ij} = -\frac{1}{l_{jj}} \sum_{k=i+1}^{n} x_{ik} l_{kj} \qquad (i > j),$$

$$x_{ij} = -\sum_{k=i+1}^{n} u_{ik} x_{kj} \qquad (i < j)$$

с такой организацией вычислений, при которой сначала вычисляется последняя строка (x_{nj}) при $j = n, n-1, \ldots, 1$, затем последний столбец (x_{in}) при $i = n-1, \ldots, 1$, потом предпоследняя строка $(x_{n-1,j})$ при $j = n-1, \ldots, 1$, и т.д.

В отличие от рассмотренной в § 2.1 схемы единственного деления схема Холецкого менее удобна для усовершенствования с целью уменьшения влияния вычислительных погрешностей путем выбора подходящих ведущих элементов. Достоинством же ее можно считать то, что LU-разложение матрицы $\mathbf{A}$ играет роль обратной матрицы, может помещаться в память компьютера на место матрицы $\mathbf{A}$ и использоваться, например, при решении нескольких систем, имеющих одну и ту же матрицу коэффициентов и разные правые части.

Замечание 2.1. Некоторое усложнение процедуры LU-факторизации позволяет применять ее в более широких условиях. А именно, для любой невырожденной квадратной матрицы $\mathbf{A}$ можно подобрать такие матрицы перестановок $\mathbf{P}$ и $\mathbf{Q}$, что будет осуществимо LU-разложение матрицы $\mathbf{PAQ}$ (причем это может быть сделано так, чтобы минимизировались вычислительные погрешности, т.е. реализовывалась стратегия выбора главного элемента). Решение системы $\mathbf{Ax} = \mathbf{b}$ в таком случае находится следующим образом. Полагаем $\mathbf{x} = \mathbf{Qz}$, где $\mathbf{z}$ — вспомогательный вектор. Тогда исходная система принимает вид $\mathbf{AQz} = \mathbf{b}$; умножив последнее равенство слева на матрицу $\mathbf{P}$, приходим к эквивалентной системе $\mathbf{PAQz} = \mathbf{Pb}$, которая, в свою очередь, может быть представлена в виде $\mathbf{LUz} = \mathbf{Pb}$. Далее последовательно решаются треугольные системы $\mathbf{Ly} = \mathbf{Pb}$ и $\mathbf{Uz} = \mathbf{y}$ относительно вспомогательных векторов $\mathbf{y}$ и $\mathbf{z}$ соответственно, после чего вычисляется искомый вектор $\mathbf{x} = \mathbf{Qz}$. Вся сложность реализации такой схемы состоит в конструировании подходящих матриц перестановок $\mathbf{P}$ и $\mathbf{Q}$.

Пример 2.2. Для матрицы $\mathbf{A}$ примера 2.1 выполнить LU-разложение и с его помощью найти второй столбец матрицы $\mathbf{A}^{-1}$.

По формулам (2.9), (2.10) последовательно вычисляем:

$u_{11} = a_{11} = 2, \quad u_{12} = a_{12} = -1, \quad u_{13} = a_{13} = 1;$

$$l_{21} = \frac{a_{21}}{u_{11}} = \frac{4}{2} = 2, \quad l_{31} = \frac{a_{31}}{u_{11}} = \frac{6}{2} = 3;$$

$u_{22} = a_{22} - l_{21} u_{12} = 3 - 2(-1) = 5, \quad u_{23} = a_{23} - l_{21} u_{13} = 1 - 2 \cdot 1 = -1;$

$$l_{32} = \frac{1}{u_{22}}(a_{32} - l_{31} u_{12}) = \frac{1}{5}(-13 - 3(-1)) = -2;$$

$u_{33} = a_{33} - l_{31} u_{13} - l_{32} u_{23} = 6 - 3 \cdot 1 - (-2)(-1) = 1.$

Таким образом, равенство $\mathbf{A} = \mathbf{LU}$ в данном случае выглядит так:

$$\begin{pmatrix} 2 & -1 & 1 \\ 4 & 3 & 1 \\ 6 & -13 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & -2 & 1 \end{pmatrix} \cdot \begin{pmatrix} 2 & -1 & 1 \\ 0 & 5 & -1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Хранить это разложение можно в виде матрицы

$$\mathbf{L} - \mathbf{E} + \mathbf{U} = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 5 & -1 \\ 3 & -2 & 1 \end{pmatrix}.$$

Чтобы найти требуемый по условию второй столбец матрицы $\mathbf{A}^{-1}$ (обозначим его $\mathbf{x}_2 = (x_{12}; x_{22}; x_{32})^T$), воспользуемся тем, что его можно считать решением векторно-матричного уравнения

$$\mathbf{A}\mathbf{x}_2 = \mathbf{e}_2, \quad \text{где} \quad \mathbf{e}_2 := (0; 1; 0)^T.$$

Наличие LU-разложения матрицы $\mathbf{A}$ позволяет свести это уравнение к двум более простым. Именно, сначала решаем уравнение $\mathbf{Ly} = \mathbf{e}_2$, т.е. систему

$$\begin{cases} y_1 = 0, \\ 2y_1 + y_2 = 1, \\ 3y_1 - 2y_2 + y_3 = 0, \end{cases}$$

получив при этом $\mathbf{y} = (y_1; y_2; y_3)^T = (0; 1; 2)^T$, а затем — аналогичное уравнение $\mathbf{U}\mathbf{x}_2 = \mathbf{y}$, также представляющее собой в развернутом виде

треугольную систему

$$\begin{cases} 2x_{12} - x_{22} + x_{32} = 0, \\ \qquad 5x_{22} - x_{32} = 1, \\ \qquad\qquad x_{32} = 2, \end{cases}$$

из которой находим $\mathbf{x}_2 = (-0.7; 0.6; 2)^T$. Следовательно, обратная к $\mathbf{A}$ матрица имеет вид

$$\mathbf{A}^{-1} = \begin{pmatrix} * & -0.7 & * \\ * & 0.6 & * \\ * & 2 & * \end{pmatrix}.$$

Заметим, что, во-первых, последние две системы можно было и не выписывать, а выполнить вычисления непосредственно по формулам (2.11), (2.13); во-вторых, перемножением элементов u_{11}, u_{22}, u_{33} легко убедиться, что при этом получается то же значение $\det \mathbf{A}$, которое было получено иначе в примере 2.1.

2.5. РАЗЛОЖЕНИЕ СИММЕТРИЧНЫХ МАТРИЦ. МЕТОД КВАДРАТНЫХ КОРНЕЙ

Объем вычислений, требующихся для решения линейных алгебраических задач с симметричными матрицами, можно сократить почти вдвое, если учитывать симметрию при треугольной факторизации матриц.

Пусть $\mathbf{A} = (a_{ij})_{i,j=1}^{n}$ — данная симметричная матрица, т.е. $a_{ij} = a_{ji}$. Будем строить ее представление в виде $\mathbf{A} = \mathbf{U}^T\mathbf{U}$, где

$$\mathbf{U} = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ 0 & u_{22} & \dots & u_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & u_{nn} \end{pmatrix}, \quad \mathbf{U}^T = \begin{pmatrix} u_{11} & 0 & \dots & 0 \\ u_{12} & u_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ u_{1n} & u_{2n} & \dots & u_{nn} \end{pmatrix}.$$

Аналогично тому, как это делалось в § 2.3, составим систему $\frac{n(n+1)}{2}$ уравнений относительно такого же количества неизвестных (элементов матрицы $\mathbf{U}$):

$$\begin{aligned} u_{11}^2 = a_{11},\ u_{12}u_{11} &= a_{12},\ \dots,\ u_{11}u_{1n} &&= a_{1n}, \\ u_{12}^2 + u_{22}^2 &= a_{22},\ \dots,\ u_{12}u_{1n} + u_{22}u_{2n} &&= a_{2n}, \\ & \dots\dots\dots\dots && \dots \\ & u_{1n}^2 + u_{2n}^2 + \dots + u_{nn}^2 &&= a_{nn}. \end{aligned}$$

Из первой строки уравнений находим сначала $u_{11} = \sqrt{a_{11}}$, затем $u_{1j} = \frac{a_{1j}}{u_{11}}$ при $j = 2, \dots, n$. Из второй — $u_{22} = \sqrt{a_{22} - u_{12}^2}$, затем $u_{2j} = \frac{a_{2j} - u_{12}u_{1j}}{u_{22}}$ при $j = 3, \dots, n$, и т.д. Завершается процесс вычислением

$$u_{nn} = \sqrt{a_{nn} - \sum_{k=1}^{n-1} u_{kn}^2}.$$

Таким образом, матрица $\mathbf{U}$ может быть определена совокупностью формул

$$\begin{aligned} u_{ii} &= \sqrt{a_{ii} - \sum_{k=1}^{i-1} u_{ki}^2} \quad \text{при} \quad i = 1, 2, \dots, n; \\ u_{ij} &= \frac{a_{ij} - \sum_{k=1}^{i-1} u_{ki}u_{kj}}{u_{ii}} \quad \text{при} \quad j = 2, \dots, n; \quad j > i \\ & (u_{ij} = 0 \quad \text{при} \quad j < i). \end{aligned} \tag{2.18}$$

Осуществимости вещественного $\mathbf{U}^T\mathbf{U}$-разложения вещественной симметричной матрицы $\mathbf{A}$ по этим формулам могут помешать два обстоятельства: обращение в нуль элемента u_{ii} при каком-либо $i \in \{1, 2, \dots, n\}$ и отрицательность подкоренного выражения. Известно, что для важного в приложениях класса симметричных положительно определенных матриц разложение по формулам (2.18) выполнимо [42, 180 и др.]*).

При наличии $\mathbf{U}^T\mathbf{U}$-разложения решение симметричной системы $\mathbf{Ax} = \mathbf{b}$ сводится к последовательному решению двух треугольных систем

$$\mathbf{U}^T\mathbf{y} = \mathbf{b} \quad \text{и} \quad \mathbf{Ux} = \mathbf{y}.$$

---

*) Более универсальным, чем $\mathbf{U}^T\mathbf{U}$-разложение Холецкого, является $\mathbf{U}^*\mathbf{D}\mathbf{U}$-разложение, пригодное для эрмитовых матриц, частным случаем которых являются симметричные (см., например, [12, 42, 158]).

Первая из них имеет вид

$$\begin{cases} u_{11}y_1 & = b_1, \\ u_{12}y_1 + u_{22}y_2 & = b_2, \\ \dots\dots\dots\dots\dots & \dots \\ u_{1n}y_1 + u_{2n}y_2 + \dots + u_{nn}y_n & = b_n, \end{cases}$$

откуда получаем вспомогательные неизвестные $y_1, y_2, \dots, y_n$ по формуле

$$y_i = \frac{b_i - \sum_{k=1}^{i-1} u_{ki} y_k}{u_{ii}}, \qquad (2.19)$$

полагая в ней $i = 1, 2, \dots, n$. Из второй системы

$$\begin{cases} u_{11}x_1 + u_{12}x_2 + \dots + u_{1n}x_n & = y_1, \\ \qquad u_{22}x_2 + \dots + u_{2n}x_n & = y_2, \\ \qquad \dots\dots\dots\dots & \\ \qquad\qquad u_{nn}x_n & = y_n \end{cases}$$

находим искомые значения x_i в обратном порядке, т.е. при $i = n, n-1, \dots, 1$, по формуле

$$x_i = \frac{y_i - \sum_{k=i+1}^{n} u_{ik} x_k}{u_{ii}}. \qquad (2.20)$$

Решение симметричных СЛАУ по формулам (2.18)–(2.20) называют ***методом квадратных корней*** или ***схемой Холецкого.*** В случае систем с положительно определенными матрицами можно ожидать хороших результатов применения такого метода (особенно, если в процессе решения делать проверку на немалость $|u_{ii}|$, чтобы избежать большого роста погрешностей). В противном случае нет, например, гарантий, что в процессе разложения не появятся чисто мнимые числа, что кстати может не отразиться на результатах, если в алгоритме реализации метода квадратных корней предусмотреть возможность появления мнимых чисел [180].

2.6. МЕТОД ПРОГОНКИ РЕШЕНИЯ СИСТЕМ С ТРЕХДИАГОНАЛЬНЫМИ МАТРИЦАМИ КОЭФФИЦИЕНТОВ

Часто возникает необходимость в решении линейных алгебраических систем, матрицы которых, являясь слабо заполненными, т.е. содержащими немного ненулевых элементов, имеют определенную структуру. Среди таких систем выделим системы с матрицами ленточной структуры, в которых ненулевые элементы располагаются на главной диагонали и на нескольких побочных диагоналях. Для решения систем с ленточными матрицами коэффициентов метод Гаусса можно трансформировать в более эффективные методы.

Рассмотрим наиболее простой случай ***ленточных систем***, к которым, как увидим впоследствии, сводится решение задач сплайн-интерполяции функций, дискретизации краевых задач для дифференциальных уравнений методами конечных разностей, конечных элементов и др. А именно, будем искать решение такой системы, каждое уравнение которой связывает три «соседних» неизвестных:

$$b_i x_{i-1} + c_i x_i + d_i x_{i+1} = r_i, \qquad (2.21)$$

где $i = 1, 2, \dots, n$; $b_1 = 0$, $d_n = 0$. Такие уравнения называют ***трехточечными разностными уравнениями второго порядка****). Система (2.21) имеет трехдиагональную структуру, что хорошо видно из следующего, эквивалентного (2.21), векторно-матричного представления:

$$\begin{bmatrix} c_1 & d_1 & 0 & 0 & \dots & 0 & 0 & 0 \\ b_2 & c_2 & d_2 & 0 & \dots & 0 & 0 & 0 \\ 0 & b_3 & c_3 & d_3 & \dots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot \\ 0 & 0 & 0 & 0 & \dots & b_{n-1} & c_{n-1} & d_{n-1} \\ 0 & 0 & 0 & 0 & \dots & 0 & b_n & c_n \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ r_{n-1} \\ r_n \end{bmatrix}.$$

---

*) Чаще вместо принятой здесь записи системы (2.21), предполагающей как бы наличие фиктивных неизвестных x_0 и x_{n+1} с нулевыми коэффициентами, считают в (2.21) i изменяющимся от 2 до $n-1$, выделяя первое и последнее уравнения системы соответственно

$$c_1 x_1 + d_1 x_2 = r_1 \quad \text{и} \quad b_n x_{n-1} + c_n x_n = r_n$$

в отдельные строки (так называемые ***краевые условия*** разностного уравнения). О разностных уравнениях см. далее § 16.3.

Преследуя, как и в § 2.1, цель избавиться от ненулевых элементов в поддиагональной части матрицы системы, предположим, что существуют такие наборы чисел δ_i и λ_i $(i = 1, 2, \ldots, n)$, при которых

$$x_i = \delta_i x_{i+1} + \lambda_i, \tag{2.22}$$

т.е. трехточечное уравнение второго порядка (2.21) преобразуется в двухточечное уравнение первого порядка (2.22). Уменьшим в связи (2.22) индекс на единицу и полученное выражение $x_{i-1} = \delta_{i-1}x_i + \lambda_{i-1}$ подставим в данное уравнение (2.21):

$$b_i\delta_{i-1}x_i + b_i\lambda_{i-1} + c_i x_i + d_i x_{i+1} = r_i,$$

откуда получаем

$$x_i = -\frac{d_i}{c_i + b_i\delta_{i-1}}x_{i+1} + \frac{r_i - b_i\lambda_{i-1}}{c_i + b_i\delta_{i-1}}.$$

Последнее равенство имеет вид (2.22) и будет точно с ним совпадать, иначе говоря, представление (2.22) будет иметь место, если при всех $i = 1, 2, \ldots, n$ выполняются рекуррентные соотношения

$$\delta_i = -\frac{d_i}{c_i + b_i\delta_{i-1}}, \quad \lambda_i = \frac{r_i - b_i\lambda_{i-1}}{c_i + b_i\delta_{i-1}}. \tag{2.23}$$

Легко видеть, что, в силу условия $b_1 = 0$, процесс вычисления δ_i, λ_i может быть начат со значений

$$\delta_1 = -\frac{d_1}{c_1}, \quad \lambda_1 = \frac{r_1}{c_1}$$

и продолжен далее по формулам (2.23) последовательно при $i = 2, 3, \ldots, n$, причем при $i = n$, в силу $d_n = 0$, получим $\delta_n = 0$. Следовательно, полагая в (2.22) $i = n$, будем иметь

$$x_n = \lambda_n = \frac{r_n - b_n\lambda_{n-1}}{c_n + b_n\delta_{n-1}}$$

(где λ_{n-1}, δ_{n-1} — уже известные с предыдущего шага числа). Далее по формулам (2.22) последовательно находятся $x_{n-1}, x_{n-2}, \ldots, x_1$ при $i = n-1, n-2, \ldots, 1$ соответственно.

Таким образом, решение уравнений вида (2.21) описываемым способом, называемым ***методом прогонки****), сводится к

*) Термин, характерный, в основном, для отечественной литературы по вычислительной математике, введен в 50-х годах XX века (см., например, [3, 55, 111]).

вычислениям по трем простым формулам: нахождение так называемых ***прогоночных коэффициентов*** δ_i, λ_i по формулам (2.23) при $i = 1, 2, \ldots, n$ (***прямая прогонка***) и затем получение неизвестных x_i по формуле (2.22) при $i = n, n-1, \ldots, 1$ (***обратная прогонка***).

Для успешного применения метода прогонки нужно, чтобы в процессе вычислений не возникало ситуаций с делением на нуль, а при больших размерностях систем не должно быть быстрого роста погрешностей округлений.

Будем называть прогонку ***корректной***, *если знаменатели прогоночных коэффициентов* (2.23) *не обращаются в нуль, и* ***устойчивой***, *если* $|\delta_i| < 1$ *при всех* $i \in \{1, 2, \ldots, n\}$.

Приведем простые достаточные условия корректности и устойчивости прогонки, которые во многих приложениях метода автоматически выполняются.

Теорема 2.2. *Пусть коэффициенты* b_i *и* d_i *уравнения* (2.21) *при* $i = 2, 3, \ldots, n-1$ *отличны от нуля и пусть*

$$|c_i| > |b_i| + |d_i| \qquad \forall\ i = 1, 2, \ldots, n. \tag{2.24}$$

Тогда прогонка (2.23), (2.22) *корректна и устойчива* (*т.е.* $c_i + b_i\delta_{i-1} \neq 0$, $|\delta_i| < 1$).

Доказательство. Воспользуемся методом математической индукции для установления обоих нужных неравенств одновременно.

При $i = 1$, в силу (2.24), имеем:

$$|c_1| > |d_1| \geq 0$$

– неравенство нулю знаменателя первой пары прогоночных коэффициентов, а также

$$|\delta_1| = \left|-\frac{d_1}{c_1}\right| < 1.$$

Предположим, что знаменатель $(i-1)$-х прогоночных коэффициентов не равен нулю и что $|\delta_{i-1}| < 1$. Тогда, используя свойства модулей, условия теоремы и индукционные предположения, получаем:

$$|c_i + b_i\delta_{i-1}| \geq |c_i| - |b_i\delta_{i-1}| > |b_i| + |d_i| - |b_i|\cdot|\delta_{i-1}| =$$
$$= |d_i| + |b_i|(1 - |\delta_{i-1}|) > |d_i| > 0,$$

а с учетом этого

$$|\delta_i| = \left|-\frac{d_i}{c_i + b_i\delta_{i-1}}\right| = \frac{|d_i|}{|c_i + b_i\delta_{i-1}|} < \frac{|d_i|}{|d_i|} = 1.$$

Следовательно, $c_i + b_i\delta_{i-1} \neq 0$ и $|\delta_i| < 1$ при всех $i \in \{1, 2, \ldots, n\}$, т.е. имеет место утверждаемая в данных условиях корректность и устойчивость прогонки. Теорема доказана.

Пусть $\mathbf{A}$ — матрица коэффициентов данной системы (2.21), удовлетворяющих условиям теоремы 2.2, и пусть

$$\delta_1 = -\frac{d_1}{c_1},\ \delta_i = -\frac{d_i}{c_i + b_i\delta_{i-1}} \quad (i = 2, 3, \ldots, n-1),\ \delta_n = 0$$

— прогоночные коэффициенты, определяемые первой из формул (2.23), а

$$\Delta_i := c_i + b_i\delta_{i-1} \quad (i = 2, 3, \ldots, n)$$

— знаменатели этих коэффициентов (отличные от нуля согласно утверждению теоремы 2.2). Непосредственной проверкой легко убедиться, что имеет место представление $\mathbf{A} = \mathbf{LU}$, где

$$\mathbf{L} = \begin{bmatrix} c_1 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ b_2 & \Delta_2 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & b_3 & \Delta_3 & 0 & \ldots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \ldots & b_{n-1} & \Delta_{n-1} & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & b_n & \Delta_n \end{bmatrix},$$

$$\mathbf{U} = \begin{bmatrix} 1 & -\delta_1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 1 & -\delta_2 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & 1 & -\delta_3 & \ldots & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \ldots & 0 & 1 & -\delta_{n-1} \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 1 \end{bmatrix},$$

единственное в силу утверждения теоремы 2.1. Как видим, LU-разложение трехдиагональной матрицы $\mathbf{A}$ может быть выполнено очень простым алгоритмом, вычисляющим Δ_i и δ_i при возрастающих значениях i. При необходимости попутно может быть вычислен определитель:

$$\det \mathbf{A} = c_1 \prod_{i=2}^{n} \Delta_i .$$

Пример 2.3. Методом прогонки решить систему

$$\begin{cases} 2x_1 + x_2 = -10, \\ 2x_1 + 9x_2 + 2x_3 = -26, \\ 4x_2 + 17x_3 - 4x_4 = -16, \\ 4x_3 + 15x_4 - 8x_5 = -2, \\ 2x_4 + 3x_5 = 16 \end{cases}$$

и вычислить определитель матрицы ее коэффициентов.

Данная система состоит из уравнений вида (2.21), где b_i, c_i, d_i, r_i суть элементы векторов

$$\mathbf{b} = \begin{pmatrix} 0 \\ 2 \\ 4 \\ 4 \\ 2 \end{pmatrix},\quad \mathbf{c} = \begin{pmatrix} 2 \\ 9 \\ 17 \\ 15 \\ 3 \end{pmatrix},\quad \mathbf{d} = \begin{pmatrix} 1 \\ 2 \\ -4 \\ -8 \\ 0 \end{pmatrix} \quad \text{и} \quad \mathbf{r} = \begin{pmatrix} -10 \\ -26 \\ -16 \\ -2 \\ 16 \end{pmatrix}$$

соответственно. Налицо диагональное преобладание в матрице системы, означающее, что ни при каком $i \in \{1, 2, \ldots, 5\}$ величина

$$\Delta_i = c_i + b_i\delta_{i-1}$$

не обратится в нуль (корректность) и что при вычислении прогоночных коэффициентов δ_i, λ_i по формулам

$$\delta_i = -\frac{d_i}{\Delta_i},\quad \lambda_i = \frac{r_i - b_i\lambda_{i-1}}{\Delta_i} \quad (i = 1, 2, \ldots, 5)$$

абсолютные величины всех δ_i окажутся меньшими единицы (устойчивость).

Полагая последовательно $i = 1, 2, \ldots, 5$, имеем:

$$\Delta_1 = c_1 = 2, \qquad \delta_1 = -\frac{d_1}{\Delta_1} = -\frac{1}{2}, \qquad \lambda_1 = \frac{r_1}{\Delta_1} = \frac{-10}{2} = -5;$$

$$\Delta_2 = c_2 + b_2\delta_1 = 8, \qquad \delta_2 = -\frac{d_2}{\Delta_2} = -\frac{1}{4}, \qquad \lambda_2 = \frac{r_2 - b_2\lambda_1}{\Delta_2} = -2;$$

$$\Delta_3 = c_3 + b_3\delta_2 = 16, \qquad \delta_3 = -\frac{d_3}{\Delta_3} = \frac{1}{4}, \qquad \lambda_3 = \frac{r_3 - b_3\lambda_2}{\Delta_3} = -\frac{1}{2};$$

$$\Delta_4 = c_4 + b_4\delta_3 = 16, \qquad \delta_4 = -\frac{d_4}{\Delta_4} = \frac{1}{2}, \qquad \lambda_4 = \frac{r_4 - b_4\lambda_3}{\Delta_4} = 0;$$

$$\Delta_5 = c_5 + b_5\delta_4 = 4, \qquad \delta_5 = -\frac{d_5}{\Delta_5} = 0, \qquad \lambda_5 = \frac{r_5 - b_5\lambda_4}{\Delta_5} = 4.$$

Обратная прогонка по формуле (2.22) при $i = 5, 4, \ldots, 1$ дает искомые

значения неизвестных:

$$x_5 = \lambda_5 = 4,$$

$$x_4 = \delta_4 x_5 + \lambda_4 = \frac{1}{2} \cdot 4 + 0 = 2,$$

$$x_3 = \delta_3 x_4 + \lambda_3 = \frac{1}{4} \cdot 2 + (-\frac{1}{2}) = 0,$$

$$x_2 = \delta_2 x_3 + \lambda_2 = -\frac{1}{4} \cdot 0 + (-2) = -2,$$

$$x_1 = \delta_1 x_2 + \lambda_1 = -\frac{1}{2} \cdot (-2) + (-5) = -4.$$

Для вычисления определителя матрицы $\mathbf{A}$ коэффициентов данной системы, как показано выше, достаточно перемножить пять чисел:

$$\det \mathbf{A} = \prod_{i=1}^{5} \Delta_i = 2 \cdot 8 \cdot 16 \cdot 16 \cdot 4 = 16384.$$

В заключение этого параграфа заметим, что, во-первых, имеются более слабые условия корректности и устойчивости прогонки, чем требуемое в теореме 2.2 условие строгого диагонального преобладания в матрице $\mathbf{A}$ (см., например, [44, 111, 158, 161]). Во-вторых, применяется ряд других, отличных от рассмотренной нами правой прогонки методов подобного типа, решающих как поставленную здесь задачу (2.21) для систем с трехдиагональными матрицами (левая прогонка, встречная прогонка, немонотонная, циклическая, ортогональная прогонки и т.д.), так и для более сложных систем с матрицами ленточной структуры или блочно-матричной структуры (например, матричная прогонка). Выводы и исследование различных вариантов метода прогонки можно найти, например, в [76, 161].

2.7. МЕТОД ВРАЩЕНИЙ РЕШЕНИЯ ЛИНЕЙНЫХ СИСТЕМ

Вернемся к рассмотрению линейных систем общего вида (2.1).

Предположим, что методом Гаусса решается система $\mathbf{Ax} = \mathbf{b}$ с матрицей коэффициентов

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{pmatrix}.$$

Приведение такой системы*) к треугольному виду прямым ходом метода Гаусса равносильно следующей последовательности эквивалентных преобразований матрицы $\mathbf{A}$:

$$\mathbf{A} \sim \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & -1 & 1 & 2 \\ 0 & -1 & -1 & 2 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & -1 & 4 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 8 \end{pmatrix}.$$

Очевидно, в случае $n \times n$-матрицы такого типа прямой ход метода Гаусса допускает рост элементов матрицы до величины 2^{n-1}. При больших n это может привести если не к переполнению разрядной сетки компьютера, то к сильному влиянию погрешностей округлений, причем в данной ситуации не даст эффекта и постолбцовый выбор главного элемента.

Рассмотренный пример оправдывает поиск других подходов к построению прямых методов решения линейных систем (2.1), возможно, более сложных, чем метод Гаусса, но не допускающих большого роста элементов в процессе преобразований и как следствие численно более устойчивых.

Как и в методе Гаусса, цель прямого хода преобразований в новом методе — приведение системы к треугольному виду последовательным обнулением поддиагональных элементов сначала первого столбца, затем второго и т.д. Делается это следующим образом.

Пусть c_1 и s_1 — некоторые отличные от нуля числа. Умножим первое уравнение системы (2.1) на c_1, второе — на s_1 и сложим их; полученным уравнением заменим первое уравнение системы. Затем первое уравнение исходной системы умножаем на $-s_1$, второе — на c_1 и результатом их сложения заменяем второе уравнение. Таким образом, первые два уравнения системы (2.1) заменяются уравнениями

$$(c_1 a_{11} + s_1 a_{21}) x_1 + (c_1 a_{12} + s_1 a_{22}) x_2 + \ldots + (c_1 a_{1n} + s_1 a_{2n}) x_n =$$
$$= c_1 b_1 + s_1 b_2 \;,$$

$$(-s_1 a_{11} + c_1 a_{21}) x_1 + (-s_1 a_{12} + c_1 a_{22}) x_2 + \ldots + (-s_1 a_{1n} + c_1 a_{2n}) x_n =$$
$$= -s_1 b_1 + c_1 b_2 \,.$$

*) Пример заимствован из [74].

На введенные два параметра c_1 и s_1 наложим два условия:

$$-s_1 a_{11} + c_1 a_{21} = 0$$

— условие обнуления (т.е. исключение x_1 из второго уравнения) и

$$c_1^2 + s_1^2 = 1$$

— условие нормировки. Легко проверить, что за c_1 и s_1, удовлетворяющие этим условиям, можно принять соответственно

$$c_1 = \frac{a_{11}}{\sqrt{a_{11}^2 + a_{21}^2}}, \quad s_1 = \frac{a_{21}}{\sqrt{a_{11}^2 + a_{21}^2}}. \qquad (2.25)$$

Эти числа можно интерпретировать как косинус и синус некото рого угла α_1 (отсюда название ***метод вращений***, так как оди промежуточный шаг прямого хода такого метода может рассмат риваться как преобразование вращения на угол α_1 расширенно матрицы системы в плоскости, определяемой индексами обну ляемого элемента*)).

После фиксирования c_1 и s_1 способом (2.25) система (2.1 принимает вид

$$\begin{cases} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \ldots + a_{1n}^{(1)}x_n = b_1^{(1)}, \\ \qquad\quad a_{22}^{(1)}x_2 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)}, \\ a_{31}x_1 + a_{32}x_2 + \ldots + a_{3n}x_n = b_3, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n, \end{cases} \qquad (2.26)$$

где

$$a_{1j}^{(1)} = c_1 a_{1j} + s_1 a_{2j} \quad (j = 1, 2, \ldots, n), \qquad b_1^{(1)} = c_1 b_1 + s_1 b_2;$$

$$a_{2j}^{(1)} = -s_1 a_{1j} + c_1 a_{2j} \quad (j = 2, 3, \ldots, n), \qquad b_2^{(1)} = -s_1 b_1 + c_1 b_2.$$

Далее первое уравнение системы (2.26) заменяется новым, полу ченным сложением результатов умножения первого и третьег

*) Более подробно эта интерпретация, как и вообще идея ортого нальных преобразований, будет рассмотрена позже применительно к ре шению полной алгебраической проблемы собственных значений (см. §§ 4.4, 4.6). Там же, в замечании 4.10 описан другой способ численно устойчивого решения СЛАУ — метод отражений.

уравнений (2.26) соответственно на

$$c_2 = \frac{a_{11}^{(1)}}{\sqrt{\left(a_{11}^{(1)}\right)^2 + a_{31}^2}} \quad \text{и} \quad s_2 = \frac{a_{31}}{\sqrt{\left(a_{11}^{(1)}\right)^2 + a_{31}^2}},$$

а третье — уравнением, полученным сложением результатов умножения тех же уравнений соответственно на $-s_2$ и c_2. Получаем систему

$$\begin{cases} a_{11}^{(2)}x_1 + a_{12}^{(2)}x_2 + \ldots + a_{1n}^{(2)}x_n = b_1^{(2)}, \\ \qquad\quad a_{22}^{(1)}x_2 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)}, \\ \qquad\quad a_{32}^{(1)}x_2 + \ldots + a_{3n}^{(1)}x_n = b_3^{(1)}, \\ a_{41}x_1 + a_{42}x_2 + \ldots + a_{4n}x_n = b_4, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n, \end{cases}$$

где

$$a_{1j}^{(1)} = c_2 a_{1j}^{(1)} + s_2 a_{3j} \quad (j = 1, 2, \ldots, n), \qquad b_1^{(1)} = c_2 b_1^{(1)} + s_2 b_3;$$

$$a_{3j}^{(1)} = -s_2 a_{1j}^{(1)} + c_2 a_{3j} \quad (j = 2, 3, \ldots, n), \qquad b_3^{(1)} = -s_2 b_1^{(1)} + c_2 b_3.$$

Проделав такие преобразования $n-1$ раз, придем к системе

$$\begin{cases} a_{11}^{(n-1)}x_1 + a_{12}^{(n-1)}x_2 + \ldots + a_{1n}^{(n-1)}x_n = b_1^{(n-1)}, \\ \qquad\quad a_{22}^{(1)}x_2 + \ldots + a_{2n}^{(1)}x_n = b_2^{(1)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \qquad\quad a_{n2}^{(1)}x_2 + \ldots + a_{nn}^{(1)}x_n = b_n^{(1)} \end{cases} \qquad (2.27)$$

такого же вида, какой приняла система (2.1) после первого этапа преобразований прямого хода метода Гаусса. Однако в отличие от (2.2) система (2.27) обладает замечательным свойством: длина любого вектора-столбца (иначе, евклидова норма) расширенной матрицы системы (2.27) остается такой же, как у соответствующего столбца исходной системы (2.1)*). Чтобы убедиться в этом,

*) Разумеется, в предположении точной реализации метода, в частности, формул типа (2.25).

достаточно посмотреть на результат следующих преобразований, учитывающих условия нормировки:

$$\left(a_{1j}^{(1)}\right)^2+\left(a_{2j}^{(1)}\right)^2=c_1^2a_{1j}^2+2c_1s_1a_{1j}a_{2j}+s_1^2a_{2j}^2+s_1^2a_{1j}^2-2c_1s_1a_{1j}a_{2j}+$$

$$+c_1^2a_{2j}^2=(c_1^2+s_1^2)a_{1j}^2+(c_1^2+s_1^2)a_{2j}^2=a_{1j}^2+a_{2j}^2.$$

Оно показывает сохранение величины суммы квадратов изменяемых на данном промежуточном шаге преобразований пары элементов любого столбца; так как остальные элементы столбцов при этом остаются неизменными, то, значит, на любом этапе преобразований длина столбца будет одной и той же, т.е. не будет наблюдаться роста элементов (но какой ценой это достигается — коэффициенты первого уравнения пересчитываются $n-1$ раз!).

Дальше точно так же за $n-2$ промежуточных шага преобразуем подсистему

$$\begin{cases} a_{22}^{(1)}x_2+\ldots+a_{2n}^{(1)}x_n=b_2^{(1)}, \\ \cdots\cdots\cdots\cdots\cdots \\ a_{n2}^{(1)}x_2+\ldots+a_{nn}^{(1)}x_n=b_n^{(1)} \end{cases}$$

системы (2.27), создавая нули под элементом $a_{22}^{(1)}$, и т.д.

В результате $n-1$ таких этапов прямого хода исходная система (2.1) будет приведена к треугольному виду (ср. с (2.3) в § 2.1):

$$\begin{cases} a_{11}^{(n-1)}x_1+a_{12}^{(n-1)}x_2+\ldots+a_{1n}^{(n-1)}x_n = b_1^{(n-1)}, \\ \qquad a_{22}^{(n-1)}x_2+\ldots+a_{2n}^{(n-1)}x_n = b_2^{(n-1)}, \\ \qquad\qquad \cdots\cdots\cdots \\ \qquad\qquad\qquad a_{nn}^{(n-1)}x_n = b_n^{(n-1)}. \end{cases}$$

Нахождение отсюда неизвестных $x_n, x_{n-1}, \ldots, x_1$ не отличается от рассмотренного ранее обратного хода метода Гаусса.

2.8. ДВА ЗАМЕЧАНИЯ К ПРИМЕНЕНИЮ ПРЯМЫХ МЕТОДОВ

1. О КОНТРОЛЕ ТОЧНОСТИ И УТОЧНЕНИИ ПРИБЛИЖЕННОГО РЕШЕНИЯ В РАМКАХ ПРЯМОГО МЕТОДА

Как отмечалось в § 2.0, прямые методы приводят к точному решению СЛАУ при точном выполнении предусматриваемых соответствующими алгоритмами арифметических операций (без округлений). Реальные же вычисления базируются на арифметике машинных (т.е. усеченных до определенного количества разрядов) чисел. Как отражается на результате решения системы подмена арифметики действительных чисел машинной арифметикой, зависит от самой решаемой системы, параметров применяемого компьютера и системы представления данных, способов реализации алгоритмов (см. по этому поводу, в частности, гл. 1). В любом случае, практически вместо точного решения СЛАУ (2.1) прямой метод дает п р и б л и ж е н н о е   р е ш е н и е*) (обозначим его $\mathbf{x}^{(0)}$). Подставив $\mathbf{x}^{(0)}$ в выражение

$$\xi := \mathbf{b}-\mathbf{A}\mathbf{x},$$

называемое ***невязкой***, по малости полученного вектора-значения $\xi^{(0)}=\mathbf{b}-\mathbf{A}\mathbf{x}^{(0)}$ можно с осторожностью судить о близости найденного решения $\mathbf{x}^{(0)}$ к точному решению $\mathbf{x}$. Если, например, $\left\|\xi^{(0)}\right\|$ — недостаточно малая величина, то следует искать вектор-***поправку*** $\mathbf{p}$ такой, что $\mathbf{x}^{(0)}+\mathbf{p}=\mathbf{x}$ есть точное решение системы (2.1), т.е.

$$\mathbf{A}\left(\mathbf{x}^{(0)}+\mathbf{p}\right)=\mathbf{b}.$$

Последнее равносильно векторно-матричному уравнению

$$\mathbf{A}\mathbf{p}=\xi^{(0)}.$$

Как видим, нахождение поправки сводится к решению такой же системы, как и (2.1), где в качестве вектора свободных членов должен быть взят вектор невязок. Поскольку матрица коэффициентов осталась той же, что и у исходной системы, нет надобности в выполнении прямого хода преобразований коэффициентов при

*) Например, метод Гаусса порождает относительную ошибку решения порядка $O\left(n\cdot 2^{-l}\cdot \operatorname{cond}\mathbf{A}\right)$, где l — число двоичных разрядов, выделяемых компьютером под мантиссу вещественного числа [158].

неизвестных (иначе, LU-разложения); достаточно выполнить только действия, касающиеся новых свободных членов (решить две треугольные системы: $\mathbf{Lz} = \xi^{(0)}$ и $\mathbf{Up} = \mathbf{z}$). Прибавив найденную поправку $\mathbf{p} = \mathbf{p}^{(0)}$ к $\mathbf{x}^{(0)}$, получаем уточненное приближенное решение $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \mathbf{p}^{(0)}$. В случае, если величина $\|\mathbf{p}^{(0)}\|$ (или $\|\mathbf{p}^{(0)}\| / \|\mathbf{x}^{(1)}\|$, если контролируется относительная, а не абсолютная погрешность) окажется недостаточно малой, процесс уточнения может быть повторен: ищется поправка $\mathbf{p}^{(1)}$ как приближенное решение уравнения $\mathbf{Ap} = \xi^{(1)}$, где $\xi^{(1)} = \mathbf{b} - \mathbf{Ax}^{(1)}$; тогда более точным должно быть решение $\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \mathbf{p}^{(1)}$. Как аргументировано утверждается в [42], сходимость к нулю невязок в таком процессе уточнения решения может не наблюдаться*), т.е. следить нужно за установлением знаков самого решения. Обычно делают не более двух-трех шагов уточнения, причем рекомендуется производить вычисление невязок в режиме накопления. Если в этом процессе не происходит сближения $\mathbf{x}^{(k)}$ при $k = 2, 3$, то это говорит скорее всего о том, что данная система плохо обусловлена и ее решение не может быть найдено с требуемой точностью без привлечения дополнительной информации об исходной задаче. В таких случаях закономерно ставить вопрос о том, что понимать под точным решением системы и, возможно, обращаться к методам нахождения ее псевдорешений (достаточно глубокие исследования затронутых здесь вопросов, основанные на изучении поведения методов при введении ошибок в исходные данные, а также ряд других сведений, в частности, точностные характеристики прямых методов, можно найти в [42]).

Хотя описанный здесь контроль точности по невязкам и уточнение решений не требует больших вычислительных затрат, требуемая память компьютера должна быть увеличена вдвое, так как при этом нужно удерживать в памяти исходные данные.

---

*) См. также пример поведения невязок у плохо обусловленной системы на приближённых решениях в гл. 1 (§§ 1.4, 1.5).

2. О ВЫЧИСЛИТЕЛЬНЫХ ЗАТРАТАХ

Одним из факторов, предопределяющих выбор того или иного метода при решении конкретной задачи, является вычислительная эффективность метода. Особенностью прямых методов является то, что здесь можно точно подсчитать требуемое количество арифметических операций. Приведем пример такого подсчета для метода прогонки решения n-мерной системы с трехдиагональной матрицей коэффициентов (см. § 2.6). Необходимые операции и их число наглядно видны из табл. 2.1 (где вычитание отождествляется со сложением):

Таблица 2.1

Подсчет арифметической сложности метода прогонки

Расчетные формулы метода прогонки	Умножений	Делений	Сложений
$\Delta_i = c_i + b_i\delta_{i-1} \quad (i = 2, \ldots, n)$	$n-1$		$n-1$
$\delta_1 = -\frac{d_1}{c_1}$; $\delta_i = -\frac{d_i}{\Delta_i} \quad (i = 2, \ldots, n-1)$		$n-1$	
$\lambda_1 = \frac{r_1}{c_1}$; $\lambda_i = \frac{r_i - b_i\lambda_{i-1}}{\Delta_i} \quad (i = 2, \ldots, n)$	$n-1$	n	$n-1$
$x_n := \lambda_n$; $\quad x_i = \delta_i x_{i+1} + \lambda_i (i = n-1, \ldots, 1)$	$n-1$		$n-1$
Итого арифметических действий	$3(n-1) + 2n - 1 + 3(n-1)$		

Учитывая, что часто операции сложения выполняются намного быстрее, чем умножения и деления, обычно ограничиваются подсчетом последних. Так, аналогично показанному нетрудно проверить, что для решения n-мерной СЛАУ методом Гаусса (без выбора главного элемента) требуется $\frac{n^3}{3} + n^2 - \frac{n}{3}$ умножений и делений*), а методом квадратных корней — $\frac{n^3}{6} + \frac{3}{2}n^2 + \frac{n}{3}$ плюс n операций извлечения корня (см. [20, 158,

---

*) Несложно подсчитать и общее число операций, включая сложение [43, 44, 61], а также время решения n-мерной системы, если известна скорость выполнения различных операций [29].

180 и др.]). Метод вращений предполагает вчетверо больше операций умножения, чем метод Гаусса [74]. При больши́х значениях размерности n существенным является старший член выражения для подсчета числа арифметических операций (иначе, *арифметической сложности* метода). Можно сказать, что вычислительные затраты на операции умножения и деления в методе Гаусса составляют величину $O\left(\frac{n^3}{3}\right)$, в методе квадратных корней $O\left(\frac{n^3}{6}\right)$, в методе вращений $O\left(\frac{4}{3}n^3\right)$, в то время как прогонка требует всего $O(5n)$ таких операций.

УПРАЖНЕНИЯ

2.1. А) Решите систему

$$\begin{cases} 0.1x_1 + 2x_2 - 10x_3 = 0.6, \\ 0.3x_1 + 6.01x_2 - 25x_3 = 1.852, \\ 0.4x_1 + 8.06x_2 + 10.001x_3 = 2.91201 \end{cases}$$

методом Гаусса, пошагово выполняя предписания алгоритма из § 2.1.

Б) Перемножением ведущих элементов метода Гаусса найдите детерминант матрицы коэффициентов данной системы.

В) Выполните задания А, Б, имитируя работу модельного компьютера, в котором под запись мантиссы числа в режиме с плавающей запятой выделяется три десятичных разряда.

Проделайте то же, проводя частичное упорядочивание по столбцам.

Сравните результаты В с результатами А и Б.

Г) Подсчитав невязки приближенных решений данной системы, полученных в задании В, произведите итерационное уточнение этих решений в той же вычислительной среде (см. § 2.8.1).

2.2. А) Выполните LU-разложение матрицы

$$\mathbf{A} = \begin{pmatrix} 4 & -3 & 2 \\ 8 & -8 & 7 \\ 12 & -5 & 5 \end{pmatrix}$$

по формулам (2.9)–(2.10) и найдите решение системы $\mathbf{Ax} = \mathbf{b}$, где $\mathbf{b} = (0; -12; 4)^T$.

Б) Используя LU-разложение, полученное в задании А, найдите матрицу $\mathbf{A}^{-1}$ двумя способами: решая подсистемы $\mathbf{Ax}_i = \mathbf{e}_i$, где $\mathbf{x}_i$ и $\mathbf{e}_i$ — столбцы соответственно искомой и единичной матриц, и реализуя вычисления по формулам (2.15)–(2.17).

2.3. А) Методом квадратных корней решите систему

$$\begin{cases} 16x_1 - 8x_2 - 4x_3 = -8, \\ -8x_1 + 13x_2 - 4x_3 - 3x_4 = 7, \\ -4x_1 - 4x_2 + 9x_3 = 6, \\ \quad - 3x_2 + 3x_4 = -3. \end{cases}$$

Б) С помощью полученного в а) $\mathbf{U}^T\mathbf{U}$-разложения Холецкого найдите детерминант матрицы коэффициентов данной системы и обратную ей матрицу.

В) Подсчитав число обусловленности, выясните, какую относительную погрешность может иметь результат А, если в одной компоненте правой части данной системы допустить абсолютную ошибку 0.01.

2.4. Выведите формулы для вычисления элементов матрицы $\mathbf{A}^{-1}$, обратной к симметричной матрице $\mathbf{A}$, на основе $\mathbf{U}^T\mathbf{U}$-разложения (подобные формулам (2.15)–(2.17)). Используйте их для обращения матрицы

$$\begin{pmatrix} 4 & 1 & -2 \\ 1 & 8 & 3 \\ -2 & 3 & 10 \end{pmatrix}.$$

2.5. Ограничиваясь трехмерным случаем, покажите, что вещественная симметричная матрица $\mathbf{A}$ может быть представлена в виде произведения $\mathbf{U}^T\mathbf{DU}$ трех вещественных матриц, где $\mathbf{U}$ — некоторая верхняя треугольная матрица, а $\mathbf{D}$ — диагональная матрица с элементами диагонали $+1$ или -1.

Примените полученный результат к системе

$$\begin{cases} x_1 + 2x_2 + 3x_3 = -3, \\ 2x_1 + x_2 + 4x_3 = -5, \\ 3x_1 + 4x_2 + x_3 = 5. \end{cases}$$

2.6. Установите, при каких $n \in \mathbf{N}$ можно гарантировать корректность и устойчивость метода прогонки для решения системы (2.21), где:

$b_i = 1 + i$ при $i = 2, 3, \ldots, n$;

$c_i = 15 + i$ при $i = 1, 2, \ldots, n$;

$d_i = -i$ при $i = 1, 2, \ldots, n-1$?

2.7. Выведите формулы л е в о й прогонки для решения системы (2.21) (т.е. такие, при которых неизвестные x_i вычислялись бы в порядке возрастания индексов).

Решите систему предыдущего упражнения при $n=7$, $r_i=i^2+14i-1$ (где $i=1, \ldots, 6$), $r_7=202$ по формулам левой и правой прогонок.

2.8. Пусть в (2.21) $b_{i+1}=d_i$ при всех $i=1, 2, \ldots, n-1$. Запишите для этого случая расчетные формулы метода квадратных корней. Подсчитайте число требуемых арифметических операций и сравните его с аналогичным результатом для метода прогонки (см. § 2.8.2).

2.9. Решите систему из упр. 2.1 методом вращений:

а) используя всю разрядную сетку калькулятора или компьютера;

б) работая, как и в упр. 2.1В, с тремя значащими цифрами.

Проанализируйте результаты, сравнивая их с результатами упражнений 2.1А и 2.1В.

ГЛАВА 3 ИТЕРАЦИОННЫЕ МЕТОДЫ РЕШЕНИЯ ЛИНЕЙНЫХ АЛГЕБРАИЧЕСКИХ СИСТЕМ И ОБРАЩЕНИЯ МАТРИЦ

Рассматриваются итерационные способы решения систем линейных алгебраических уравнений и обращения матриц, служащие серьезной альтернативой прямым методам решения таких задач, по крайней мере в случаях, когда их размерность велика. Показывается логика построения нескольких наиболее важных итерационных процессов, таких как методы простых итераций, Якоби, Зейделя, релаксации, Шульца, и изучаются условия сходимости последовательностей приближений, получаемых этими методами, к искомым решениям. Дается первое представление о методах установления. Приводится алгоритм метода сопряженных градиентов и объясняется суть метода минимальных невязок.

3.1. РЕШЕНИЕ СЛАУ МЕТОДОМ ПРОСТЫХ ИТЕРАЦИЙ

Система стандартного вида

$$\mathbf{Ax}=\mathbf{b}, \tag{3.1}$$

где $\mathbf{A}=(a_{ij})_{i,\,j=1}^{n}$ — $n\times n$-матрица, а $\mathbf{x}=(x_1,\ldots,x_n)^T$ и $\mathbf{b}=(b_1,\ldots,b_n)^T$ — n-мерные векторы-столбцы, тем или иным способом (таких способов существует бесконечное множество; некоторые из них будут рассмотрены ниже) может быть преобразована к эквивалентной ей системе вида

$$\mathbf{x}=\mathbf{Bx}+\mathbf{c}, \tag{3.2}$$

где $\mathbf{x}$ — тот же вектор неизвестных, а $\mathbf{B}$ и $\mathbf{c}$ — некоторые новые матрица и вектор соответственно. Систему (3.2) можно трактовать как задачу о неподвижной точке линейного отображения $\mathbf{B}$ в пространстве $\mathbf{R}_n$ и по аналогии со скалярным случаем (более подробно изучаемым в гл.6) определить последовательность приближений $\mathbf{x}^{(k)}$ к неподвижной точке $\mathbf{x}^*$ рекуррентным равенством

$$\mathbf{x}^{(k+1)}=\mathbf{Bx}^{(k)}+\mathbf{c}, \quad k=0,1,2,\ldots. \tag{3.3}$$

Итерационный процесс (3.3), начинающийся с некоторого вектора $\mathbf{x}^{(0)} = \left(x_1^{(0)}, \ldots, x_n^{(0)}\right)^T$, будем называть ***методом простых итераций*** (коротко МПИ).

Изучим комплекс вопросов о сходимости этого процесса. А именно: 1) какие нужно предъявить требования к $\mathbf{B}$, $\mathbf{c}$ и $\mathbf{x}^{(0)}$, чтобы последовательность $\left(\mathbf{x}^{(k)}\right)$ при $k \to \infty$ имела пределом $\mathbf{x}^*$ — неподвижную точку задачи (3.2) (и значит, решение эквивалентной (3.2) исходной системы (3.1))? 2) с какой скоростью сходится этот процесс, т.е. каков закон убывания абсолютных погрешностей получаемых по формуле (3.3) приближений? 3) сколько нужно сделать итераций по формуле (3.3), чтобы при заданном начальном приближении $\mathbf{x}^{(0)}$ найти решение задачи (3.2) с заданной точностью?

Ответы на подобные вопросы теории итерационных методов в $\mathbf{R}_n$ часто опираются на следующие два утверждения о сходимости степенных матричных рядов, точнее, «матричной геометрической прогрессии». Во втором из них, а также всюду далее под нормой матрицы понимается мультипликативная норма такая, что $\|\mathbf{E}\| = 1$ ($\mathbf{E}$ — единичная матрица).

Лемма 3.1*). *Условие, что все собственные числа матрицы* $\mathbf{B}$ *по модулю меньше* 1, *является необходимым и достаточным для того, чтобы*:

1) $\mathbf{B}^k \to \mathbf{0}$ *при* $k \to \infty$ $(k \in \mathbf{N})$;

2) *матрица* $\mathbf{E}-\mathbf{B}$ *имела обратную и*

$$(\mathbf{E}-\mathbf{B})^{-1} = \mathbf{E}+\mathbf{B}+\mathbf{B}^2+\ldots+\mathbf{B}^k+\ldots .$$

Лемма 3.2**). *Если* $\|\mathbf{B}\| \le q < 1$, *то матрица* $\mathbf{E}-\mathbf{B}$ *имеет обратную матрицу* $(\mathbf{E}-\mathbf{B})^{-1} = \sum_{k=0}^{\infty}\mathbf{B}^k$ *и при этом справедливо неравенство*

$$\left\|(\mathbf{E}-\mathbf{B})^{-1}\right\| \le \frac{1}{1-q}.$$

---

*) В некоторых литературных источниках лемма 3.1 называется ***леммой Неймана*** (см., например, [139]).

) В функциональном анализе для более общего случая, когда $\mathbf{B}$ — линейный оператор, действующий в полных нормированных пространствах, эту лемму называют ***теоремой Банаха* ([80] и др.).

Доказательство. Рассмотрим матричный ряд

$$\mathbf{E}+\mathbf{B}+\mathbf{B}^2+\ldots+\mathbf{B}^k+\ldots \qquad (3.4)$$

В силу условия леммы и вытекающего из мультипликативного свойства нормы неравенства $\left\|\mathbf{B}^k\right\| \le \|\mathbf{B}\|^k$, этот ряд можно промажорировать сходящимся числовым рядом:

$$\|\mathbf{E}\|+\|\mathbf{B}\|+\left\|\mathbf{B}^2\right\|+\ldots+\left\|\mathbf{B}^k\right\|+\ldots \le 1+q+q^2+\ldots+q^k+\ldots = \frac{1}{1-q}.$$

Следовательно, ряд (3.4) сходится, т.е. существует матрица

$$\mathbf{V} = \mathbf{E}+\mathbf{B}+\mathbf{B}^2+\ldots+\mathbf{B}^k+\ldots$$

такая, что $\|\mathbf{V}\| \le \dfrac{1}{1-q}$. Так как

$$(\mathbf{E}-\mathbf{B})\mathbf{V} = (\mathbf{E}-\mathbf{B})\left(\mathbf{E}+\mathbf{B}+\mathbf{B}^2+\ldots+\mathbf{B}^k+\ldots\right) =$$
$$= \mathbf{E}+\mathbf{B}+\mathbf{B}^2+\ldots+\mathbf{B}^k+\ldots-\mathbf{B}-\mathbf{B}^2-\mathbf{B}^3-\ldots-\mathbf{B}^{k+1}-\ldots = \mathbf{E},$$

то $\mathbf{V} = (\mathbf{E}-\mathbf{B})^{-1}$. Лемма доказана.

Доказательство леммы 1 более сложно. Его можно найти во многих учебных пособиях по вычислительной математике и по функциональному анализу (см., например, [20, 61, 80, 99, 139]).

Теорема 3.1. *Необходимым и достаточным условием сходимости метода простых итераций* (3.3) *при любом начальном векторе* $\mathbf{x}^{(0)}$ *к решению* $\mathbf{x}^*$ *системы* (3.2) *является требование, чтобы все собственные числа матрицы* $\mathbf{B}$ *были по модулю меньше* 1.

Доказательство. *Достаточность.* Пусть $\max|\lambda_B| < 1$, тогда по лемме 3.1 общий член $\mathbf{B}^k$ ряда (3.4) стремится к нуль-матрице и существует матрица $(\mathbf{E}-\mathbf{B})^{-1}$, являющаяся пределом частичных сумм $\left(\mathbf{E}+\mathbf{B}+\mathbf{B}^2+\ldots+\mathbf{B}^k\right)$ при $k \to \infty$. Применяя рекурсию в равенстве (3.3), определяющем МПИ, получим:

$$\mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)}+\mathbf{c} = \mathbf{B}^2\mathbf{x}^{(k-1)}+(\mathbf{B}+\mathbf{E})\mathbf{c} = \ldots =$$
$$= \mathbf{B}^{k+1}\mathbf{x}^{(0)}+\left(\mathbf{E}+\mathbf{B}+\mathbf{B}^2+\ldots+\mathbf{B}^k\right)\mathbf{c}. \qquad (3.5)$$

В силу сказанного выше, предел последнего выражения существует при любом фиксированном $\mathbf{x}^{(0)}$ и равен $(\mathbf{E}-\mathbf{B})^{-1}\mathbf{c}$. Следо-

вательно, итерационный процесс (3.3) сходится и

$$\mathbf{x}^* := \lim_{k\to\infty} \mathbf{x}^{(k)} = (\mathbf{E}-\mathbf{B})^{-1}\mathbf{c}.$$

Подставляя $\mathbf{x}^*$ в уравнение (3.2), преобразованное к виду $(\mathbf{E}-\mathbf{B})\mathbf{x} = \mathbf{c}$, имеем:

$$(\mathbf{E}-\mathbf{B})(\mathbf{E}-\mathbf{B})^{-1}\mathbf{c} = \mathbf{c},$$

т.е. вектор $\mathbf{x}^* = \lim_{k\to\infty} \mathbf{x}^{(k)}$ удовлетворяет системе (3.2). (Заметим, что это $\mathbf{x}^*$ — единственное решение (3.2). Действительно, допустив, что наряду с $\mathbf{x}^*$ таким, что $\mathbf{x}^* = \mathbf{B}\mathbf{x}^* + \mathbf{c}$, имеется $\mathbf{x}^{**}$, удовлетворяющее такому же равенству $\mathbf{x}^{**} = \mathbf{B}\mathbf{x}^{**} + \mathbf{c}$, получаем $\mathbf{x}^* - \mathbf{x}^{**} = \mathbf{B}(\mathbf{x}^* - \mathbf{x}^{**})$. Последнее означает, что число $\lambda = 1$ по определению является собственным значением матрицы $\mathbf{B}$, что противоречит условию).

Необходимость. Как видно из представления общего члена итерационной последовательности $(\mathbf{x}^{(k)})$ в форме (3.5), существование $\lim_{k\to\infty} \mathbf{x}^{(k+1)}$ при любых векторах $\mathbf{x}^{(0)}$ и $\mathbf{c}$ (в том числе и нулевых, что гарантирует существование предела каждого слагаемого в правой части (3.5)) влечет сходимость матриц $\mathbf{B}^{k+1}$ к нуль-матрице и сходимость ряда $\sum_{k=0}^{\infty}\mathbf{B}^k$ к $(\mathbf{E}-\mathbf{B})^{-1}$. Согласно лемме 3.1, это равносильно выполнению условия $|\lambda_B| < 1$ для каждого собственного числа матрицы $\mathbf{B}$.

Теорема доказана.

Теорема 3.2. *Пусть* $\|\mathbf{B}\| \le q < 1$. *Тогда при любом начальном векторе* $\mathbf{x}^{(0)}$ *МПИ* (3.3) *сходится к единственному решению* $\mathbf{x}^*$ *задачи* (3.2) *и при всех* $k \in \mathbf{N}$ *справедливы оценки погрешности*:

1) $\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \dfrac{q}{1-q}\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\|$ (***апостериорная***);

2) $\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \dfrac{q^k}{1-q}\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|$ (***априорная***).*)

*) Лат. *a priori* и *a posteriori* означают соответственно «до опыта» и «из опыта», т.е. априорной оценкой можно воспользоваться до начала счета, а апостериорной — лишь после проведения k-й итерации.

(Одно и то же обозначение $\|\cdot\|$ здесь используется для матричных и векторных норм, согласованных между собой, т.е. таких, что $\|\mathbf{A}\mathbf{x}\| \le \|\mathbf{A}\|\cdot\|\mathbf{x}\|$).

Доказательство. Вычитая из равенства (3.3) равенство $\mathbf{x}^{(k)} = \mathbf{B}\mathbf{x}^{(k-1)} + \mathbf{c}$, имеем $\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} = \mathbf{B}\left(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right)$. Переходя в последнем к нормам, получаем неравенство

$$\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \le q\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\|, \qquad (3.6)$$

из которого видно, в силу условия $q < 1$, что элементы итерационной последовательности $(\mathbf{x}^{(k)})$ сближаются с ростом номера k. С помощью (3.6) оценим разность между $(k+m)$-м и k-м членами этой последовательности при некотором $m \in \mathbf{N}$:

$$\begin{aligned}
&\left\|\mathbf{x}^{(k+m)} - \mathbf{x}^{(k)}\right\| = \left\|\mathbf{x}^{(k+m)} - \mathbf{x}^{(k+m-1)} + \mathbf{x}^{(k+m-1)} - \mathbf{x}^{(k+m-2)} +\right.\\
&\quad \left. + \mathbf{x}^{(k+m-2)} - \ldots - \mathbf{x}^{(k+1)} + \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \le\\
&\le \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| + \left\|\mathbf{x}^{(k+2)} - \mathbf{x}^{(k+1)}\right\| + \ldots + \left\|\mathbf{x}^{(k+m)} - \mathbf{x}^{(k+m-1)}\right\| \le\\
&\le q\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| + q^2\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| + \ldots + q^m\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| =\\
&= \frac{q(1-q^m)}{1-q}\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| = \frac{q^k}{1-q}(1-q^m)\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|.
\end{aligned}$$

Рассматривая итоговое неравенство при $k \to \infty$ и фиксированном m, видим, что $(\mathbf{x}^{(k)})$ является фундаментальной последовательностью и, в силу полноты пространства $\mathbf{R}_n$, имеет предел. Обозначим его $\mathbf{x}^*$. Переходя к пределу в равенстве (3.3), получаем равенство $\mathbf{x}^* = \mathbf{B}\mathbf{x}^* + \mathbf{c}$, означающее, что $\mathbf{x}^* = \lim_{k\to\infty} \mathbf{x}^{(k)}$ — решение уравнения (3.2). При этом $\mathbf{x}^*$ — единственное решение (3.2), так как предположив существование другого решения $\mathbf{x}^{**} \ne \mathbf{x}^*$ и нормируя равенство $\mathbf{x}^* - \mathbf{x}^{**} = \mathbf{B}(\mathbf{x}^* - \mathbf{x}^{**})$, приходим к противоречащему условию теоремы неравенству $\|\mathbf{B}\| \ge 1$.

Справедливость утверждаемых в теореме оценок погрешности видна из неравенств

$$\left\|\mathbf{x}^{(k+m)} - \mathbf{x}^{(k)}\right\| \le \frac{q(1-q^m)}{1-q}\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \le \frac{q^k}{1-q}(1-q^m)\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|,$$

если в них теперь зафиксировать k и перейти к пределу при $m \to \infty$. Теорема доказана.

Замечание 3.1. Последние неравенства говорят еще о том, что априорная оценка, как правило, грубее апостериорной.

Замечание 3.2. Теорема 3.2 могла быть доказана на основе леммы 3.2 и теоремы 3.1. В частности, сходимость последовательности $(\mathbf{x}^{(k)})$ к решению $\mathbf{x}^*$ системы (3.2) сразу следует из теоремы 3.1, в силу соотношений $|\lambda_{\mathbf{B}}| \le \|\mathbf{B}\| < 1$. Из леммы 3.2 также легко вывести другую априорную оценку погрешности k-го приближения: вычитая из равенства

$$\mathbf{x}^* = (\mathbf{E}-\mathbf{B})^{-1}\mathbf{c} = (\mathbf{E}+\mathbf{B}+\ldots+\mathbf{B}^k+\ldots)\mathbf{c}$$

равенство

$$\mathbf{x}^{(k)} = \mathbf{B}^k\mathbf{x}^{(0)} + (\mathbf{E}+\mathbf{B}+\ldots+\mathbf{B}^{k-1})\mathbf{c}$$

(см. (3.5)), имеем:

$$\|\mathbf{x}^* - \mathbf{x}^{(k)}\| = \|(\mathbf{B}^k + \mathbf{B}^{k+1} + \ldots)\mathbf{c} - \mathbf{B}^k\mathbf{x}^{(0)}\| \le$$

$$\le \|\mathbf{B}^k\| \cdot \|(\mathbf{E}-\mathbf{B})^{-1}\mathbf{c} - \mathbf{x}^{(0)}\| \le q^k\left(\|\mathbf{x}^{(0)}\| + \frac{\|\mathbf{c}\|}{1-q}\right).$$

Замечание 3.3. Априорная оценка позволяет подсчитывать заранее число итераций k, достаточное для получения решения $\mathbf{x}^*$ с заданной точностью ε (в смысле допустимого уровня абсолютных погрешностей) при выбранном начальном векторе $\mathbf{x}^{(0)}$. Для этого нужно найти наименьшее целое решение неравенства

$$\frac{q^k}{1-q}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| \le \varepsilon$$

относительно переменной k (или неравенства $q^k\left(\|\mathbf{x}^{(0)}\| + \frac{\|\mathbf{c}\|}{1-q}\right) \le \varepsilon$ в соответствии с результатом предыдущего замечания). Апостериорной же оценкой удобно пользоваться непосредственно в процессе вычислений и останавливать этот процесс, как только выполнится неравенство

$$\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \le \frac{1-q}{q}\varepsilon.$$

Отметим, что неравенство $\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\| \le \varepsilon$ будет гарантией выполнения неравенства $\|\mathbf{x}^* - \mathbf{x}^{(k)}\| \le \varepsilon$ только в том случае, когда $q \le \frac{1}{2}$.

Замечание 3.4. По поводу выбора начального приближения.
Как установлено выше, сходимость МПИ (3.3) при условии $|\lambda_{\mathbf{B}}| < 1$

гарантируется при любом начальном векторе $\mathbf{x}^{(0)}$. Очевидно, итераций потребуется тем меньше, чем ближе $\mathbf{x}^{(0)}$ к $\mathbf{x}^*$. Если нет никакой дополнительной информации о решении задачи (3.2) (например, может быть известным решение близкой задачи или грубое решение данной задачи), за $\mathbf{x}^{(0)}$ обычно принимают вектор $\mathbf{c}$ свободных членов системы (3.2). Мотивация этого может быть такой: матрица $\mathbf{B}$ «мала», значит вектор $\mathbf{Bx}$ «мал», следовательно, и вектор $\mathbf{x}^*$ не должен сильно отличаться от вектора $\mathbf{c}$. При выборе $\mathbf{x}^{(0)} = \mathbf{c}$ фигурирующая в теореме 3.2 априорная оценка принимает вид

$$\|\mathbf{x}^* - \mathbf{x}^{(k)}\| \le \frac{\|\mathbf{c}\|}{1-q}q^{k+1} \quad \forall\, k \in \mathbf{N}.$$

Пример 3.1. Для системы

$$\begin{cases} 1.1x_1 - 0.2x_2 + 0.1x_3 = 1.6, \\ 0.1x_1 - 1.2x_2 - 0.2x_3 = 2.3, \\ 0.2x_1 - 0.1x_2 + 1.1x_3 = 1.5 \end{cases}$$

записать какой-нибудь сходящийся процесс простых итераций. За сколько шагов этого процесса, начатого с нуль-вектора, можно гарантированно достичь точности $\varepsilon = 0.001$ по норме-максимум? Найти третье приближение, оценить его абсолютную погрешность и сравнить ее с истинной погрешностью, зная точное решение системы $\mathbf{x}^* = (1; -2; 1)^T$.

Учитывая очевидную близость матрицы данной системы к единичной матрице, вычленим единицы из диагональных элементов, в результате чего система преобразуется к виду

$$\begin{cases} x_1 = -0.1x_1 + 0.2x_2 - 0.1x_3 + 1.6, \\ x_2 = 0.1x_1 - 0.2x_2 - 0.2x_3 - 2.3, \\ x_3 = -0.2x_1 + 0.1x_2 - 0.1x_3 + 1.5. \end{cases}$$

Эта система равносильна исходной и имеет форму уравнения (3.2), в котором можно считать

$$\mathbf{B} = \begin{pmatrix} -0.1 & 0.2 & -0.1 \\ 0.1 & -0.2 & -0.2 \\ -0.2 & 0.1 & -0.1 \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} 1.6 \\ -2.3 \\ 1.5 \end{pmatrix}.$$

Так как $\|\mathbf{B}\|_\infty = 0.5\ (=: q) < 1$, можно воспользоваться теоремой 3.2. Согласно ей, метод простых итераций

$$\begin{cases} x_1^{(k+1)} = -0.1x_1^{(k)} + 0.2x_2^{(k)} - 0.1x_3^{(k)} + 1.6, \\ x_2^{(k+1)} = 0.1x_1^{(k)} - 0.2x_2^{(k)} - 0.2x_3^{(k)} - 2.3, \\ x_3^{(k+1)} = -0.2x_1^{(k)} + 0.1x_2^{(k)} - 0.1x_3^{(k)} + 1.5 \end{cases}$$

(где $k = 0, 1, 2, \ldots$) определяет сходящуюся к решению $\mathbf{x}^*$ последовательность векторов $\mathbf{x}^{(k)} = \left(x_1^{(k)};\ x_2^{(k)};\ x_3^{(k)}\right)^T$, априорная оценка погрешностей которых есть

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\|_\infty \le \frac{0.5^k}{1-0.5}\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|_\infty .$$

При заданном векторе $\mathbf{x}^{(0)} = \mathbf{0}$ первым приближением $\mathbf{x}^{(1)}$, очевидно, служит вектор $\mathbf{c}$ свободных членов и, значит, $\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|_\infty = \|\mathbf{c}\|_\infty = 2.3$. Следовательно, требуемое число итерационных шагов, достаточное для достижения точности 0.001, может быть найдено как первое из последовательности натуральных чисел k, удовлетворяющих неравенству

$$0.5^{k-1} \cdot 2.3 \le 0.001,$$

т. е. получив отсюда $k \approx 12.2$, принимаем $k = 13$.

Вычислим приближения $\mathbf{x}^{(2)}$ и $\mathbf{x}^{(3)}$:

$$\begin{cases} x_1^{(2)} = -0.1\cdot 1.6 + 0.2\cdot(-2.3) - 0.1\cdot 1.5 + 1.6 = 0.83, \\ x_2^{(2)} = 0.1\cdot 1.6 - 0.2\cdot(-2.3) - 0.2\cdot 1.5 - 2.3 = -1.98, \\ x_3^{(2)} = -0.2\cdot 1.6 + 0.1\cdot(-2.3) - 0.1\cdot 1.5 + 1.5 = 0.8; \end{cases}$$

$$\begin{cases} x_1^{(3)} = -0.1\cdot 0.83 + 0.2\cdot(-1.98) - 0.1\cdot 0.8 + 1.6 = 1.041, \\ x_2^{(3)} = 0.1\cdot 0.83 - 0.2\cdot(-1.98) - 0.2\cdot 0.8 - 2.3 = -1.981, \\ x_3^{(3)} = -0.2\cdot 0.83 + 0.1\cdot(-1.98) - 0.1\cdot 0.8 + 1.5 = 1.056. \end{cases}$$

Априорная оценка погрешности третьего приближения дает

$$\left\|\mathbf{x}^* - \mathbf{x}^{(3)}\right\|_\infty \le \frac{0.5^3}{1-0.5}\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|_\infty = 0.25\cdot 2.3 = 0.575,$$

в то время как истинная ошибка составляет величину

$$\left\|\mathbf{x}^* - \mathbf{x}^{(3)}\right\|_\infty = \left\|\begin{pmatrix} -0.041 \\ -0.019 \\ -0.044 \end{pmatrix}\right\|_\infty = 0.044,$$

что на порядок лучше прогнозируемой ошибки. Это говорит о том, что найденное выше априорное число итерационных шагов наверняка больше необходимого (оценка есть оценка!). Если воспользоваться апостериорной оценкой погрешности, то для того же приближения $\mathbf{x}^{(3)}$ получим

$$\left\|\mathbf{x}^* - \mathbf{x}^{(3)}\right\|_\infty \le \frac{0.5}{1-0.5}\left\|\mathbf{x}^{(3)} - \mathbf{x}^{(2)}\right\|_\infty = \left\|\begin{pmatrix} 0.211 \\ -0.001 \\ 0.256 \end{pmatrix}\right\|_\infty = 0.256$$

— несколько лучший результат, как и следовало ожидать в соответствии с замечанием 3.1. Нетрудно понять, что сравнительная точность апостериорных оценок заметно увеличивается с увеличением номера итерации.

3.2. МЕТОД ЯКОБИ

Вернемся к рассмотрению СЛАУ в виде (3.1). После выяснения условия, которому должна удовлетворять матрица коэффициентов приведенной системы (3.2) для сходимости МПИ (3.3), следует осуществить приведение системы (3.1) к виду (3.2) так, чтобы это условие выполнялось. Рассмотрим один из способов такого приведения, достаточно эффективный в определенных случаях.

Представим матрицу $\mathbf{A}$ системы (3.1) в виде

$$\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{R},$$

где $\mathbf{D}$ — диагональная, а $\mathbf{L}$ и $\mathbf{R}$ — соответственно левая и правая строго треугольные (т.е. с нулевой диагональю) матрицы. Тогда система (3.1) может быть записана в виде

$$\mathbf{Lx} + \mathbf{Dx} + \mathbf{Rx} = \mathbf{b}, \tag{3.7}$$

и если на диагонали исходной матрицы нет нулей, то эквивалентной (3.1) задачей вида (3.2) будет

$$\mathbf{x} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{R})\mathbf{x} + \mathbf{D}^{-1}\mathbf{b}, \tag{3.8}$$

т.е. в равенствах (3.2) и (3.3) следует положить

$$\mathbf{B} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{R}), \quad \mathbf{c} = \mathbf{D}^{-1}\mathbf{b}.$$

Основанный на таком приведении системы (3.1) к виду (3.2) метод простых итераций (3.3) называют ***методом Якоби****). В векторно-матричных обозначениях он определяется формулой

$$\mathbf{x}^{(k+1)} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{R})\mathbf{x}^{(k)} + \mathbf{D}^{-1}\mathbf{b}, \quad k = 0, 1, 2, \ldots. \tag{3.9}$$

Чтобы записать метод Якоби (3.9) решения системы (3.1) в развернутом виде, достаточно заметить, что обратной матрицей к матрице $\mathbf{D} = (a_{ii})_{i=1}^n$ служит диагональная матрица $\mathbf{D}^{-1}$ с элементами диагонали $d_{ii} = 1/a_{ii}$. Поэтому представление (3.8) систе-

*) Якоби Карл Густав Якоб (1804–1851) — немецкий математик.

мы (3.1), записанной в виде (3.7), равнозначно выражению «диагональных неизвестных» через остальные:

$$\begin{cases} x_1 = -(a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n - b_1)/a_{11}, \\ x_2 = -(a_{21}x_1 + a_{23}x_3 + \ldots + a_{2n}x_n - b_2)/a_{22}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n = -\left(a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{n,n-1}x_{n-1} - b_n\right)/a_{nn}. \end{cases} \qquad (3.10)$$

Теперь для записи итерационного процесса (3.9) осталось в равенствах системы (3.10) только «навесить» индексы, соответствующие номерам приближений (т.е. $k = 0, 1, 2, \ldots$):

$$\begin{cases} x_1^{(k+1)} = -\left(a_{12}x_2^{(k)} + a_{13}x_3^{(k)} + \ldots + a_{1n}x_n^{(k)} - b_1\right)/a_{11}, \\ x_2^{(k+1)} = -\left(a_{21}x_1^{(k)} + a_{23}x_3^{(k)} + \ldots + a_{2n}x_n^{(k)} - b_2\right)/a_{22}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^{(k+1)} = -\left(a_{n1}x_1^{(k)} + a_{n2}x_2^{(k)} + \ldots + a_{n,n-1}x_{n-1}^{(k)} - b_n\right)/a_{nn}. \end{cases} \qquad (3.9а)$$

Установим простой достаточный признак сходимости метода Якоби к решению системы (3.1).

Теорема 3.3. *В случае диагонального преобладания в матрице* **A** *системы* (3.1) *метод Якоби* (3.9) *сходится.*

Доказательство. ***Диагональное преобладание*** в матрице **A** означает, что

$$|a_{ii}| > \sum_{\substack{j=1 \\ (j \neq i)}}^{n} |a_{ij}| \qquad \forall\, i = 1, \ldots, n.$$

Следовательно, в любой строке ***матрицы итерирования*** (иначе, ***матрицы перехода***)

$$\mathbf{B} = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & -\frac{a_{13}}{a_{11}} & \ldots & -\frac{a_{1,n-1}}{a_{11}} & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & -\frac{a_{23}}{a_{22}} & \ldots & -\frac{a_{2,n-1}}{a_{22}} & -\frac{a_{2n}}{a_{22}} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & -\frac{a_{n3}}{a_{nn}} & \ldots & -\frac{a_{n,n-1}}{a_{nn}} & 0 \end{pmatrix} \qquad (3.11)$$

метода Якоби сумма модулей элементов меньше единицы. Значит, по крайней мере, одна из норм матрицы **B** (согласованная, в частности, с векторной нормой-максимум) меньше единицы. Таким образом, существуют нормы, в которых к методу Якоби, рассматриваемому как метод простых итераций, применима теорема 3.2, т.е. метод Якоби сходится. Так как сходимость по одной норме в пространстве $\mathbf{R}_n$ означает сходимость по любой другой, тем самым теорема 3.3 доказана.

Замечание 3.5. Обратим внимание на то, что к методу Якоби при условии диагонального преобладания в матрице **A** относится полностью заключение теоремы 3.2, а также предыдущие замечания; нужно лишь учесть в них, что матрица **B** определяется с помощью (3.11), а вектор **c** — равенством

$$\mathbf{c} = \left(\frac{b_1}{a_{11}}; \frac{b_2}{a_{22}}; \ldots; \frac{b_n}{a_{nn}}\right)^T.$$

При этом матрица **D** в представлении системы (3.7) заведомо обратима.

Следствием теоремы 1, устанавливающим необходимые и достаточные условия сходимости метода Якоби, является следующая теорема.

Теорема 3.4. *Метод Якоби* (3.9) *сходится к решению системы* (3.1) *в том и только в том случае, когда все корни уравнения*

$$\begin{vmatrix} a_{11}\lambda & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22}\lambda & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn}\lambda \end{vmatrix} = 0$$

по модулю меньше единицы.

Действительно, чтобы все собственные числа матрицы $\mathbf{B} = -\mathbf{D}^{-1}(\mathbf{L}+\mathbf{R})$ были по модулю меньше единицы, как этого требует теорема 3.1 для данного случая, нужно, чтобы меньше единицы были модули всех корней характеристического уравнения

$$\det\left(-\mathbf{D}^{-1}(\mathbf{L}+\mathbf{R}) - \lambda\mathbf{E}\right) = 0.$$

Последнее же эквивалентно уравнению

$$\det(\mathbf{L}+\mathbf{R}+\lambda\mathbf{D}) = 0,$$

которое в записи через элементы исходной матрицы **A** и фигурирует в формулировке теоремы.

3.3. МЕТОД ЗЕЙДЕЛЯ

Под ***методом Зейделя**** обычно понимается такое видоизменение метода простых итераций (3.3) решения СЛАУ, приведенных к виду (3.2), при котором для подсчета i-й компоненты $(k+1)$-го приближения к искомому вектору $\mathbf{x}^*$ используются уже найденные на этом, т.е. $(k+1)$-м шаге, новые значения первых $i-1$ компонент. Это означает, что если система (3.1) тем или иным способом сведена к системе (3.2) с матрицей коэффициентов $\mathbf{B}=(b_{ij})_{i,j=1}^n$ и вектором свободных членов $\mathbf{c}=(c_i)_{i=1}^n$, то приближения к ее решению по методу Зейделя определяются системой равенств

$$\begin{cases} x_1^{(k+1)} = b_{11}x_1^{(k)} + b_{12}x_2^{(k)} + \ldots + b_{1n}x_n^{(k)} + c_1, \\ x_2^{(k+1)} = b_{21}x_1^{(k+1)} + b_{22}x_2^{(k)} + \ldots + b_{2n}x_n^{(k)} + c_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^{(k+1)} = b_{n1}x_1^{(k+1)} + b_{n2}x_2^{(k+1)} + \ldots + b_{nn}x_n^{(k)} + c_n, \end{cases} \tag{3.12}$$

где $k=0,1,2,\ldots$, а $x_i^{(0)}$ — компоненты заданного (выбранного) начального вектора $\mathbf{x}^{(0)}$.

Применительно к компьютерным расчетам один полный, т.е. векторный итерационный шаг метода Зейделя может интерпретироваться как реализация формулы (3.2), где под знаком равенства следует понимать знак присваивания, а под $\mathbf{x}$ — один и тот же линейный массив из n элементов, на нулевом шаге заполненный компонентами заданного начального вектора $\mathbf{x}^{(0)}$. В таком случае на обычной однопроцессорной вычислительной машине элементы массива $\mathbf{x}$ будут постепенно замещаться новыми элементами.

Аналогичный взгляд на МПИ (3.3) показывает, что для компьютерной реализации одного его шага нужно целиком сохранять n-элементный массив $\mathbf{x}$, подставляемый в правую часть, до тех пор, пока не сформируется полностью другой n-элементный массив — результат данного итерационного шага. В связи с такой интерпретацией метод Зейделя называют ***методом последовательных смещений***, а метод простых итераций — ***одновременных смещений***.

---

*) Зейдель Филипп Людвиг (1821–1896) — немецкий астроном и математик.

Пример 3.2. Возьмем за основу расчетные формулы МПИ, записанные в примере 3.1 для заданной там системы, и в соответствии с (3.12) преобразуем их в расчетные формулы метода Зейделя:

$$\begin{cases} x_1^{(k+1)} = -0.1x_1^{(k)} + 0.2x_2^{(k)} - 0.1x_3^{(k)} + 1.6, \\ x_2^{(k+1)} = 0.1x_1^{(k+1)} - 0.2x_2^{(k)} - 0.2x_3^{(k)} - 2.3, \\ x_3^{(k+1)} = -0.2x_1^{(k+1)} + 0.1x_2^{(k+1)} - 0.1x_3^{(k)} + 1.5. \end{cases}$$

Начиная процесс вычислений с того же начального приближения $\mathbf{x}^{(0)} = \mathbf{0}$, далее при $k = 0, 1, 2$ последовательно получаем:

$$\begin{cases} x_1^{(1)} = 1.6, \\ x_2^{(1)} = 0.1\cdot 1.6 - 2.3 = -2.14, \\ x_3^{(1)} = -0.2\cdot 1.6 + 0.1\cdot(-2.14) + 1.5 = 0.966; \end{cases}$$

$$\begin{cases} x_1^{(2)} = -0.1\cdot 1.6 + 0.2\cdot(-2.14) - 0.1\cdot 0.966 + 1.6 \approx 0.915, \\ x_2^{(2)} = 0.1\cdot 0.915 - 0.2\cdot(-2.14) - 0.2\cdot 0.966 - 2.3 \approx -1.974, \\ x_3^{(2)} = -0.2\cdot 0.915 + 0.1\cdot(-1.974) - 0.1\cdot 0.966 + 1.5 \approx 1.023; \end{cases}$$

$$\begin{cases} x_1^{(3)} = -0.1\cdot 0.915 + 0.2\cdot(-1.974) - 0.1\cdot 1.023 + 1.6 \approx 1.011, \\ x_2^{(3)} = 0.1\cdot 1.011 - 0.2\cdot(-1.974) - 0.2\cdot 1.023 - 2.3 \approx -2.009, \\ x_3^{(3)} = -0.2\cdot 1.011 + 0.1\cdot(-2.009) - 0.1\cdot 1.023 + 1.5 \approx 0.995. \end{cases}$$

Вектор ошибок третьего приближения $\mathbf{x}^{(3)}$ по методу Зейделя к точному решению $\mathbf{x}^* = (1;\,-2;\,1)^T$ данной в примере 3.1 линейной системы есть $(-0.011;\ 0.009;\ 0.005)^T$. Его норма-максимум составляет величину 0.011, что в 4 раза меньше, чем при применении в тех же условиях метода простых итераций. Поскольку показанное улучшение не требует дополнительных вычислений, налицо эффективность подобной модификации МПИ и целесообразность ее дальнейшего изучения.

Остановимся подробнее на случае, когда приведение системы (3.1) к виду (3.2) основано на представлении (3.7), т.е. когда метод Зейделя есть модификация метода Якоби. Запись соответствующих расчетных формул здесь сводится к верхней

индексации системы (3.10) по типу (3.12):

$$\begin{cases} x_1^{(k+1)} = -\left(a_{12}x_2^{(k)} + a_{13}x_3^{(k)} + \ldots + a_{1n}x_n^{(k)} - b_1\right)/a_{11}, \\ x_2^{(k+1)} = -\left(a_{21}x_1^{(k+1)} + a_{23}x_3^{(k)} + \ldots + a_{2n}x_n^{(k)} - b_2\right)/a_{22}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^{(k+1)} = -\left(a_{n1}x_1^{(k+1)} + a_{n2}x_2^{(k+1)} + \ldots + a_{n,\,n-1}x_{n-1}^{(k+1)} - b_n\right)/a_{nn}, \end{cases} \quad (3.13)$$

где $k = 0, 1, 2, \ldots$; $\mathbf{x}^{(0)} = \left(x_1^{(0)}, \ldots, x_n^{(0)}\right)^T$ задается.

Для анализа сходимости метода Зейделя (3.13) обратимся к его векторно-матричной форме. Легко видеть, что если неявный вид метода Якоби, вытекающий из представления (3.7) системы (3.1), есть

$$\mathbf{Lx}^{(k)} + \mathbf{Dx}^{(k+1)} + \mathbf{Rx}^{(k)} = \mathbf{b} \quad \text{(сравните с (3.9))},$$

то равнозначный (3.13) неявный вид метода Зейделя в векторно-матричных обозначениях суть

$$\mathbf{Lx}^{(k+1)} + \mathbf{Dx}^{(k+1)} + \mathbf{Rx}^{(k)} = \mathbf{b}.$$

Следовательно, тот же вектор $\mathbf{x}^{(k+1)}$, который фигурирует в левой части совокупности равенств (3.13), может быть получен по формуле

$$\mathbf{x}^{(k+1)} = -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{Rx}^{(k)} + (\mathbf{L}+\mathbf{D})^{-1}\mathbf{b}. \quad (3.14)$$

Последнее выражение определяет не что иное, как МПИ (3.3) для системы вида (3.2), где

$$\mathbf{B} = -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{R}, \quad \mathbf{c} = (\mathbf{L}+\mathbf{D})^{-1}\mathbf{b},$$

т.е. результат применения одного шага метода Зейделя (3.13), полученного на основе $(\mathbf{L}+\mathbf{D}+\mathbf{R})$-разложения матрицы $\mathbf{A}$, можно расценивать как шаг МПИ для эквивалентной (3.1) задачи о неподвижной точке

$$\mathbf{x} = -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{Rx} + (\mathbf{L}+\mathbf{D})^{-1}\mathbf{b} \quad (3.15)$$

(разумеется, если треугольная матрица $\mathbf{L}+\mathbf{D}$ обратима). Эта связь между методом Зейделя и методом простых итераций позволяет легко переформулировать некоторые утверждения о сходимости МПИ применительно к методу Зейделя (3.13).

Теорема 3.5. *Для сходимости метода Зейделя* (3.13) *необходимо и достаточно, чтобы все корни уравнения*

$$\begin{vmatrix} a_{11}\lambda & a_{12} & \ldots & a_{1n} \\ a_{21}\lambda & a_{22}\lambda & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1}\lambda & a_{n2}\lambda & \ldots & a_{nn}\lambda \end{vmatrix} = 0 \quad (3.16)$$

были по модулю меньше единицы.

Доказательство. Применяя теорему 3.1 к МПИ (3.14), составляем характеристическое уравнение, определяющее собственные числа λ матрицы $\mathbf{B} = -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{R}$:

$$\det\left(-(\mathbf{L}+\mathbf{D})^{-1}\mathbf{R} - \lambda\mathbf{E}\right) = 0.$$

Это уравнение равносильно уравнению

$$\det((\mathbf{L}+\mathbf{D})\lambda + \mathbf{R}) = 0,$$

которое с учетом представления $\mathbf{A} = \mathbf{L}+\mathbf{D}+\mathbf{R}$ совпадает с (3.16).

Замечание 3.6. Уравнение (3.16), а также метод (3.13), являющийся частным случаем более общей формы метода Зейделя (3.12), называют иногда соответственно ***уравнением*** и ***методом Некрасова*** [180]. Метод (3.12) называют еще и ***методом Гаусса–Зейделя*** [138, 139].

Прямым следствием теоремы 3.2 для метода Зейделя (3.13) является следующая теорема.

Теорема 3.6. *Пусть* $\left\|(\mathbf{L}+\mathbf{D})^{-1}\mathbf{R}\right\| \le t < 1$. *Тогда при любом начальном векторе* $\mathbf{x}^{(0)}$ *метод Зейделя* (3.13) *сходится к решению* $\mathbf{x}^*$ *системы* (3.1) *и справедливы оценки погрешности*

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \frac{t}{1-t}\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \le \frac{t^k}{1-t}\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\|. \quad (3.17)$$

Ясно, что для непосредственного использования оценок (3.17) нужно предварительно выполнить обращение треугольной матрицы $\mathbf{L}+\mathbf{D}$ и перемножить матрицы $(\mathbf{L}+\mathbf{D})^{-1}$ и $\mathbf{R}$. В таком случае частично теряется смысл в поэлементной реализации метода Зейделя (3.13); вместо этого можно проводить итерации по

формуле (3.14) до тех пор, пока не выполнится условие $\left\| \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)} \right\| \le \frac{1-t}{t}\varepsilon$, где $\varepsilon > 0$ — требуемая точность. В частности, такой подход может быть рекомендован при решении СЛАУ методом Зейделя на компьютерах с векторной обработкой информации.

Более подходящие для использования оценки погрешности метода Зейделя (3.13) можно получить, разлагая матрицу $\mathbf{B} := -\mathbf{D}^{-1}(\mathbf{L}+\mathbf{R})$ (см. (3.11)) в сумму двух строго треугольных матриц, т.е. полагая

$$\mathbf{B} = \mathbf{B}_L + \mathbf{B}_R,$$

где

$$\mathbf{B}_L := \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ -\frac{a_{21}}{a_{22}} & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & \cdots & -\frac{a_{n,n-1}}{a_{nn}} & 0 \end{pmatrix},$$

$$\mathbf{B}_R := \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & \cdots & -\frac{a_{1,n-1}}{a_{11}} & -\frac{a_{1n}}{a_{11}} \\ 0 & 0 & \cdots & -\frac{a_{2,n-1}}{a_{22}} & -\frac{a_{2n}}{a_{22}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

С ними эквивалентное (3.1) уравнение (3.2) приобретает вид

$$\mathbf{x} = \mathbf{B}_L x + \mathbf{B}_R \mathbf{x} + \mathbf{c},$$

т.е. для решения $\mathbf{x}^*$ будет точным равенство

$$\mathbf{x}^* = \mathbf{B}_L \mathbf{x}^* + \mathbf{B}_R \mathbf{x}^* + \mathbf{c},$$

а метод Зейделя (3.13) — соответственно

$$\mathbf{x}^{(k+1)} = \mathbf{B}_L \mathbf{x}^{(k+1)} + \mathbf{B}_R \mathbf{x}^{(k)} + \mathbf{c}.$$

Из двух последних равенств получаем следующее:

$$\mathbf{x}^* - \mathbf{x}^{(k+1)} = \mathbf{B}_L(\mathbf{x}^* - \mathbf{x}^{(k+1)}) + \mathbf{B}_R(\mathbf{x}^* - \mathbf{x}^{(k)}).$$

Это равенство, записанное в виде

$$\mathbf{x}^* - \mathbf{x}^{(k)} = \mathbf{B}_L(\mathbf{x}^* - \mathbf{x}^{(k)}) + \mathbf{B}_R(\mathbf{x}^* - \mathbf{x}^{(k-1)}), \qquad (3.18)$$

можно расценивать как точную связь между погрешностями k-го и $(k-1)$-го приближений в методе Зейделя (3.13). Отсюда, переходя к нормам, легко вывести априорную оценку погрешности, что можно оформить в виде следующего утверждения.

Теорема 3.7 [3]. *Пусть* $\frac{\|\mathbf{B}_R\|}{1-\|\mathbf{B}_L\|} \le p < 1$. *Тогда метод Зейделя* (3.13) *определяет сходящуюся последовательность* $(x^{(k)})$ *при любом начальном векторе* $\mathbf{x}^{(0)}$, *и имеет место оценка*

$$\left\| \mathbf{x}^* - \mathbf{x}^{(k)} \right\| \le p^k \left\| \mathbf{x}^* - \mathbf{x}^{(0)} \right\| \quad \forall\, k \in \mathrm{N}.$$

Как и у предыдущей, у этой теоремы имеются свои недостатки, затрудняющие ее применение: нужно знать меру близости начального приближения $\mathbf{x}^{(0)}$ к решению $\mathbf{x}^*$. Ценность ее скорее в том, что в ней фигурирует легко вычисляемый коэффициент $\frac{\|\mathbf{B}_R\|}{1-\|\mathbf{B}_L\|}$ связи ошибок результатов двух соседних итерационных шагов, характеризующий быстроту сходимости метода Зейделя (3.13). При организации практических вычислений по формулам (3.13) целесообразнее ориентироваться на следующий результат.

Теорема 3.8. *Пусть* $\|\mathbf{B}\| < 1$ (*где* $\mathbf{B}$ — *матрица* (3.11)). *Тогда для определяемой методом Зейделя* (3.13) *последовательности приближений справедлива апостериорная оценка погрешности*

$$\left\| \mathbf{x}^* - \mathbf{x}^{(k)} \right\| \le \frac{\|\mathbf{B}_R\|}{1-\|\mathbf{B}\|} \left\| \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)} \right\| \quad \forall\, k \in \mathrm{N}.$$

Для доказательства этого утверждения подставим $\mathbf{B}_L = \mathbf{B} - \mathbf{B}_R$ в равенство (3.18). Имеем

$$\mathbf{x}^* - \mathbf{x}^{(k)} = \mathbf{B}_L(\mathbf{x}^* - \mathbf{x}^{(k)}) + \mathbf{B}_R(\mathbf{x}^* - \mathbf{x}^{(k-1)}),$$

что в условиях теоремы (с учетом леммы 3.2) равносильно равенству

$$\mathbf{x}^* - \mathbf{x}^{(k)} = (\mathbf{E} - \mathbf{B})^{-1} \mathbf{B}_R(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}).$$

Отсюда, переходя к нормам, получаем нужную оценку.

Из теоремы 3.8 вытекает следующая, более удобная на практике, формулировка.

Следствие 3.1. *Пусть k_ε — первое из натуральных чисел k, с которым при заданном $\varepsilon > 0$ для генерируемой процессом Зейделя* (3.13) *последовательности векторов* $\mathbf{x}^{(k)} = \left(x_1^{(k)}, \ldots, x_n^{(k)}\right)^T$ *в некоторых согласованных нормах выполняется неравенство*

$$\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \le \frac{1-\|\mathbf{B}\|}{\|\mathbf{B}_R\|}\varepsilon .$$

Тогда за решение $\mathbf{x}^*$ *системы* (3.1) *может быть принят вектор* $\mathbf{x}^{(k_\varepsilon)}$, *и абсолютная погрешность при этом не будет превышать* ε *(в выбранной норме).*

Условия сходимости методов Зейделя и простых итераций, вообще говоря, различаются. Но некоторые достаточные условия можно применять к обоим методам одновременно.

Теорема 3.9. *Если в матрице* **A** *системы* (3.1) *имеет место диагональное преобладание, то метод Зейделя* (3.13) *сходится, причем быстрее, чем метод Якоби* (3.9а).

Доказательство. Вычитая тождественное (3.13) равенство (3.14) из равенства (3.15), рассматриваемого как верное равенство при подстановке в него решения $\mathbf{x}^*$, получаем

$$\mathbf{x}^* - \mathbf{x}^{(k+1)} = -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{R}\left(\mathbf{x}^* - \mathbf{x}^{(k)}\right).$$

Введем в рассмотрение вектор ошибок

$$\mathbf{\Delta}^{(k)} = \mathbf{x}^* - \mathbf{x}^{(k)}$$

с компонентами $\delta_i^{(k)} = x_i^* - x_i^{(k)}$. Тогда это равенство через элементы матрицы **A** исходной системы (3.1) можно записать так (см. соответствие между (3.14) и (3.13)):

$$\delta_i^{(k+1)} = -\frac{1}{a_{ii}}\sum_{j=1}^{i-1} a_{ij}\delta_j^{(k+1)} - \frac{1}{a_{ii}}\sum_{j=i+1}^{n} a_{ij}\delta_j^{(k)},$$

где $i = 1, 2, \ldots, n$; $k = 0, 1, 2, \ldots$. Переходя к модулям, отсюда имеем:

$$\left|\delta_i^{(k+1)}\right| \le \frac{1}{|a_{ii}|}\sum_{j=1}^{i-1}|a_{ij}|\cdot\left|\delta_j^{(k+1)}\right| + \frac{1}{|a_{ii}|}\sum_{j=i+1}^{n}|a_{ij}|\cdot\left|\delta_j^{(k)}\right| \le$$

$$\le \left(\max_j\left|\delta_j^{(k+1)}\right|\right)\cdot\frac{1}{|a_{ii}|}\sum_{j=1}^{i-1}|a_{ij}| + \left(\max_j\left|\delta_j^{(k)}\right|\right)\cdot\frac{1}{|a_{ii}|}\sum_{j=i+1}^{n}|a_{ij}|.$$

Обозначим $\alpha_i = \frac{1}{|a_{ii}|}\sum_{j=1}^{i-1}|a_{ij}|$, $\beta_i = \frac{1}{|a_{ii}|}\sum_{j=i+1}^{n}|a_{ij}|$ и $\left\|\mathbf{\Delta}^{(k)}\right\|_\infty = \max_i\left|\delta_i^{(k)}\right|$ (где $\left\|\mathbf{\Delta}^{(k)}\right\|_\infty$ может трактоваться как абсолютная погрешность k-го приближения по методу Зейделя*)). В этих обозначениях последнее неравенство имеет вид

$$\left|\delta_i^{(k+1)}\right| \le \alpha_i\left\|\mathbf{\Delta}^{(k+1)}\right\|_\infty + \beta_i\left\|\mathbf{\Delta}^{(k)}\right\|_\infty. \tag{3.19}$$

Пусть $m \in \{1, 2, \ldots, n\}$ — значение индекса i, при котором реализуется равенство $\left|\delta_m^{(k+1)}\right| = \max_i\left|\delta_i^{(k+1)}\right| = \left\|\mathbf{\Delta}^{(k+1)}\right\|_\infty$. Тогда из (3.19) следует

$$\left\|\mathbf{\Delta}^{(k+1)}\right\|_\infty \le \alpha_m\left\|\mathbf{\Delta}^{(k+1)}\right\|_\infty + \beta_m\left\|\mathbf{\Delta}^{(k)}\right\|_\infty,$$

т.е. при этом фиксированном $i = m$ выполняется неравенство

$$\left\|\mathbf{\Delta}^{(k+1)}\right\|_\infty \le \frac{\beta_m}{1-\alpha_m}\left\|\mathbf{\Delta}^{(k)}\right\|_\infty. \tag{3.20}$$

Так как в условиях диагонального преобладания справедливо неравенство $\alpha_i + \beta_i < 1$, а это неравенство, в свою очередь, влечет неравенство $\frac{\beta_i}{1-\alpha_i} \le \alpha_i + \beta_i$ (проверьте!), причем равенство в последнем случае имеет место лишь при $i = 1$, то абсолютная погрешность приближений по методу Зейделя (3.13), согласно (3.20), убывает со скоростью геометрической прогрессии, знаменатель которой, вообще говоря, меньше, чем для соответствующего этому случаю метода Якоби (3.9а). Такое мажорирование последовательности величин $\left\|\mathbf{x}^* - \mathbf{x}^{(k+1)}\right\|$ позволяет сделать заключение о справедливости доказываемой теоремы.

Замечание 3.7. В соответствии с последней теоремой в *методе Зейделя* (3.13) вместо оценок (3.17), требующих дополнительных затрат на обращение треугольной матрицы, *допустимо использование оценок погрешности метода Якоби*. Естественно, они заведомо грубее.

---

*) Индекс ∞ у знака нормы использован согласно обозначению соответствующего частного случая l_p-нормы (иначе, нормы Гёльдера, см. приложение 1).

Остановимся еще на одном важном для приложений классе систем вида (3.1), для которых имеет место сходимость метода Зейделя (3.13).

Определение 3.1 [61]. *Система* $\mathbf{Ax}=\mathbf{b}$ *называется* ***нормальной****, если матрица* $\mathbf{A}$ *— симметричная положительно определенная.*

Теорема 3.10. *Если система* (3.1) *— нормальная, то метод Зейделя* (3.13) *сходится.*

Доказательство этой теоремы заключается в проверке того, что положительная определенность матрицы $\mathbf{A}=\mathbf{L}+\mathbf{D}+\mathbf{L}^T$ влечет выполнимость условия теоремы 3.5 (т.е. собственные числа матрицы $-(\mathbf{L}+\mathbf{D})^{-1}\mathbf{L}^T$ по модулю меньше единицы). Это доказательство можно найти, например, в [20, 61].

Любая линейная система $\mathbf{Ax}=\mathbf{b}$ легко может быть симметризована умножением на матрицу $\mathbf{A}^T$. Более того, справедлива следующая теорема.

Теорема 3.11 [61]. *Пусть* $\det\mathbf{A}\neq 0$. *Тогда система* $\mathbf{A}^T\mathbf{Ax}=\mathbf{A}^T\mathbf{b}$ *— нормальная.**)

Таким образом, если, например, известно, что система (3.1) однозначно разрешима, но в ее матрице коэффициентов нет диагонального преобладания, метод Зейделя типа (3.13) можно применять к системе $\mathbf{A}^T\mathbf{Ax}=\mathbf{A}^T\mathbf{b}$. Правда, здесь возникают трудности со своевременным окончанием процесса итерирования, обеспечивающим заданную точность приближенного решения, так как приведенные ранее оценки погрешности (см. теорему 3.6 и замечание 3.7) в этом случае часто «не работают». Да и сходимость при этом может оказаться весьма медленной.

Наряду с рассмотренными, применяют и другие способы приведения систем (3.1) к виду (3.2) для их решения методами простых итераций и Зейделя. Достаточно общий подход к этой процедуре заключается в том, что эквивалентное (3.1) уравнение $\mathbf{0}=\mathbf{b}-\mathbf{Ax}$ умножается на некоторую неособенную матрицу $\mathbf{H}$ (матричный параметр) и к обеим частям прибавляется вектор $\mathbf{x}$. Полученное уравнение

$$\mathbf{x}=\mathbf{x}+\mathbf{H}(\mathbf{b}-\mathbf{Ax}),$$

*) Переход от системы $\mathbf{Ax}=\mathbf{b}$ к системе $\mathbf{A}^T\mathbf{Ax}=\mathbf{A}^T\mathbf{b}$ (или в более общем случае к $\mathbf{A}^*\mathbf{Ax}=\mathbf{A}^*\mathbf{b}$) называют ***симметризацией Гаусса***.

переписанное в виде

$$\mathbf{x}=(\mathbf{E}-\mathbf{HA})\mathbf{x}+\mathbf{Hb},$$

имеет структуру (3.2). Проблема теперь заключается в подборе матрицы $\mathbf{H}$ такой, чтобы матрица $\mathbf{B}=\mathbf{E}-\mathbf{HA}$ обладала нужными свойствами для сходимости применяемых методов; для некоторых классов матриц $\mathbf{A}$ имеются определенные рекомендации [61, 99]. Заметим, что матрица $\mathbf{H}$ может быть как постоянной (в этом случае говорят о ***стационарном*** итерационном процессе), так и изменяющейся от шага к шагу. В последнем случае данное уравнение $\mathbf{Ax}=\mathbf{b}$ подменяется последовательностью эквивалентных ему задач $\mathbf{x}=\mathbf{B}_k\mathbf{x}+\mathbf{c}_k$, и соответствующий итерационный процесс называется ***нестационарным***.

3.4. ПОНЯТИЕ О МЕТОДЕ РЕЛАКСАЦИИ

В случаях, когда применение оценок погрешностей в методах простых итераций и Зейделя невозможно из-за отсутствия констант $q<1$ или $t<1$, ограничивающих сверху какие-либо нормы матрицы итерирования соответствующего метода (см. теоремы 3.2 и 3.6), эти методы неэффективны и, более того, как будет показано в § 3.7, малонадежны ввиду медленной сходимости. Рассмотрим одно обобщение метода Зейделя, позволяющее иногда в несколько раз ускорить сходимость итерационной последовательности.

Пусть $z_i^{(k)}$ — обозначение i-й компоненты k-го приближения к решению системы (3.1) по методу Зейделя, а обозначение $x_i^{(k)}$ будем использовать для i-й компоненты k-го приближения, получаемого новым методом. Этот метод определим равенством

$$x_i^{(k+1)}=x_i^{(k)}+\omega\left(z_i^{(k+1)}-x_i^{(k)}\right), \qquad (3.21)$$

где $i=1,2,\ldots,n$; $k=0,1,2,\ldots$; $x_i^{(0)}$ — задаваемые начальные значения; ω — числовой параметр, который называют ***параметром релаксации***. Очевидно, при $\omega=1$ метод (3.21), называемый ***методом релаксации*** (*ослабления*), совпадает с методом Зейделя*).

Конкретизируем метод релаксации для случая, когда исходная система (3.1) представляется в виде (3.7) и, следовательно, метод Зейделя имеет вид (3.13).

*) Метод Зейделя в качестве представителя семейства релаксационных методов называют иногда ***методом полной релаксации***.

Пользуясь введенными здесь обозначениями, запишем на основании (3.13) дополнительное к (3.21) равенство для выражения компонент векторов $\mathbf{z}^{(k)}=\left(z_i^{(k)}\right)_{i=1}^{n}$ через компоненты векторов $\mathbf{x}^{(k)}=\left(x_i^{(k)}\right)_{i=1}^{n}$:

$$z_i^{(k+1)}=\frac{1}{a_{ii}}\left(b_i-\sum_{j=1}^{i-1}a_{ij}x_j^{(k+1)}-\sum_{j=i+1}^{n}a_{ij}x_j^{(k)}\right). \qquad (3.22)$$

Таким образом, метод релаксации можно понимать как поочередное применение формул (3.22) и (3.21) при каждом $k=0,1,2,\ldots$. Действительно, задав начальные значения $x_1^{(0)},x_2^{(0)},\ldots,x_n^{(0)}$ и параметр ω, при $k=0$, полагая $i=1,2,\ldots,n$, вычислим

$$z_1^{(1)},x_1^{(1)};\quad z_2^{(1)},x_2^{(1)};\quad\ldots;\quad z_n^{(1)},x_n^{(1)};$$

при $k=1$, так же полагая $i=1,2,\ldots,n$, находим

$$z_1^{(2)},x_1^{(2)};\quad z_2^{(2)},x_2^{(2)};\quad\ldots;\quad z_n^{(2)},x_n^{(2)}$$

и т.д. Но можно избавиться от вспомогательной последовательности $\left(\mathbf{z}^{(k)}\right)$, подставив (3.22) в (3.21). Для $i=1,2,\ldots,n$ будем иметь:

$$x_i^{(k+1)}=(1-\omega)x_i^{(k)}+\frac{\omega}{a_{ii}}\left(b_i-\sum_{j=1}^{i-1}a_{ij}x_j^{(k+1)}-\sum_{j=i+1}^{n}a_{ij}x_j^{(k)}\right). \qquad (3.23)$$

От формулы (3.23), объединяющей формулы (3.22) и (3.21) и пригодной для проведения покоординатных вычислений, мало отличающихся от вычислений по методу Зейделя, легко перейти к векторно-матричной записи процесса релаксации. С этой целью перепишем (3.23) в виде

$$a_{ii}x_i^{(k+1)}+\omega\sum_{j=1}^{i-1}a_{ij}x_j^{(k+1)}=(1-\omega)a_{ii}x_i^{(k)}-\omega\sum_{j=i+1}^{n}a_{ij}x_j^{(k)}+\omega b_i$$

и далее, учитывая аддитивное представление матрицы $\mathbf{A}=\mathbf{L}+\mathbf{D}+\mathbf{R}$, получаем векторно-матричный итерационный процесс в неявной форме

$$(\mathbf{D}+\omega\mathbf{L})\mathbf{x}^{(k+1)}=(1-\omega)\mathbf{D}\mathbf{x}^{(k)}-\omega\mathbf{R}\mathbf{x}^{(k)}+\omega\mathbf{b}.$$

Умножив последнее равенство слева на матрицу $(\mathbf{D}+\omega\mathbf{L})^{-1}$,

приходим к эквивалентному (3.23) методу простых итераций

$$\mathbf{x}^{(k+1)}=(\mathbf{D}+\omega\mathbf{L})^{-1}((1-\omega)\mathbf{D}-\omega\mathbf{R})\mathbf{x}^{(k)}+\omega(\mathbf{D}+\omega\mathbf{L})^{-1}\mathbf{b} \qquad (3.24)$$

(подстановка сюда значения $\omega=1$ превращает (3.24) в МПИ (3.14), эквивалентный методу Зейделя (3.13)).

Представление метода релаксации (3.23) в виде (3.24) позволяет сделать для него некоторые утверждения о сходимости, на основании соответствующих теорем о сходимости МПИ. Например, можно применить теоремы 3.1 и 3.2, полагая в них $\mathbf{B}=(\mathbf{D}+\omega\mathbf{L})^{-1}((1-\omega)\mathbf{D}-\omega\mathbf{R})$, правда, получаемые при этом результаты вряд ли будут вызывать интерес. Более глубокие результаты на этом пути получают, изучая спектральные свойства таких матриц $\mathbf{B}$. Так, установлено, что *для сходимости процесса* (3.23) *необходимо, чтобы* $\omega\in(0,2)$. Для некоторых классов СЛАУ (3.1) это требование к параметру релаксации является и достаточным. Справедлива следующая теорема, обобщающая теорему 3.8.

Теорема 3.12 (***Островского–Рейча*** [138, 161]). *Для нормальной системы* $\mathbf{A}\mathbf{x}=\mathbf{b}$ *метод релаксации* (3.23) *сходится при любом* $\mathbf{x}^{(0)}$ *и любом* $\omega\in(0,2)$.

Поскольку итерационный процесс (3.23) содержит параметр, естественно распорядиться им так, чтобы сходимость последовательности $\left(\mathbf{x}^{(k)}\right)$ была наиболее быстрой. Очевидно, это достигается в том случае, когда спектральный радиус матрицы $\mathbf{B}=(\mathbf{D}+\omega\mathbf{L})^{-1}((1-\omega)\mathbf{D}-\omega\mathbf{R})$ будет минимальным. В общем случае задача нахождения оптимального значения $\omega=\omega_0$ не решена, и в практических расчетах применяют метод проб и ошибок. Однако для отдельных важных классов задач такие значения удается выразить через собственные числа матрицы $\mathbf{D}^{-1}(\mathbf{L}+\mathbf{R})$ (т.е. корни уравнения, фигурирующего в теореме 3.4) и даже оценить ускорение, достигаемое введением в процесс Зейделя оптимального параметра релаксации. Существенно отметить, что это оптимальное значение $\omega_0\in(1,2)$. При значениях $\omega\in(1,2)$ метод (3.23) называют ***методом последовательной верхней релаксации*** (сокращенно ПВР- или SOR-методом*)). Ввиду низкой эффективности метода (3.23) при $\omega\in(0,1)$, называемого в этом случае ***методом нижней релаксации***, название «метод ПВР» в последнее время относят ко всему семейству методов (3.23), т.е.

*) От англ. *Successive over relaxation.*

для любых $\omega \in (0, 2)$. При этом случай $\omega \in (1, 2)$ называют ***сверхрелаксацией***.

Покажем возможный выигрыш при использовании метода ПВР на простейшем примере.

Пример 3.3. Для системы

$$\begin{cases} 2x_1 + x_2 = 1, \\ x_1 + 2x_2 = -1 \end{cases}$$

с симметричной положительно определенной матрицей $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ и очевидным решением $\mathbf{x}^* = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ выполним по три итерационных шага, начиная с $\mathbf{x}^{(0)} = \mathbf{0}$, методами Якоби, Зейделя и ПВР соответственно по формулам

$$\begin{cases} x_1^{(k+1)} = -0.5x_2^{(k)} + 0.5, \\ x_2^{(k+1)} = -0.5x_1^{(k)} - 0.5, \end{cases} \qquad \begin{cases} x_1^{(k+1)} = -0.5x_2^{(k)} + 0.5, \\ x_2^{(k+1)} = -0.5x_1^{(k+1)} - 0.5 \end{cases}$$

и

$$\begin{cases} x_1^{(k+1)} = (1-\omega)x_1^{(k)} + \frac{\omega}{2}(1 - x_2^{(k)}), \\ x_2^{(k+1)} = (1-\omega)x_2^{(k)} - \frac{\omega}{2}(1 + x_1^{(k+1)}) \end{cases} \quad \text{при } \omega = 1.1.$$

Сравнительные результаты третьего шага представлены следующей таблицей.

Таблица 3.1

	Метод Якоби	Метод Зейделя	Метод ПВР (с $\omega = 1.1$)
$x_1^{(3)}$	0.875	≈ 0.969	≈ 1.0008
$x_2^{(3)}$	–0.875	≈ -0.984	≈ -1.0009
$\lVert x^* - x^{(3)} \rVert_\infty$	0.125	≈ 0.031	< 0.001

Значение параметра релаксации ω здесь взято близким к оптимальному, которое для матриц «*упорядоченных согласованно со свойством* A» [27, 138] находится по формуле

$$\omega_{опт} = \frac{2}{1 + \sqrt{1 - \rho^2(\mathbf{B})}},$$

где $\rho(\mathbf{B})$ — спектральный радиус матрицы $\mathbf{B} = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{R})$ (в данном

случае $\mathbf{B} = \begin{pmatrix} 0 & 0.5 \\ 0.5 & 0 \end{pmatrix}$, $\rho(\mathbf{B}) = 0.5$, $\omega_{опт} \approx 1.0718$).

3.5. О ДРУГИХ ИТЕРАЦИОННЫХ МЕТОДАХ РЕШЕНИЯ СЛАУ

В основе построения и изучения или, по крайней мере, понимания многих итерационных методов лежит связь между системами алгебраических уравнений и методами дискретизации дифференциальных уравнений, их порождающими.

В простейшем абстрактном, но далеко не самом общем случае, легко установить такую связь между СЛАУ (3.1) и абстрактным дифференциальным уравнением

$$\frac{d\mathbf{y}}{dt} + \mathbf{A}\mathbf{y}(t) = \mathbf{b} \qquad (3.25)$$

с начальным условием $\mathbf{y}(0) = \mathbf{x}^{(0)}$, где t — абстрактная скалярная переменная, изменяющаяся на промежутке $[0, +\infty)$, а матрица $\mathbf{A}$ и вектор $\mathbf{b}$ те же, что и в уравнении (3.1).

Пусть постоянный вектор $\mathbf{x}$ и переменный вектор $\mathbf{y} = \mathbf{y}(t)$ — решения задач (3.1) и (3.25) соответственно. Введем вектор

$$\mathbf{z}(t) := \mathbf{x} - \mathbf{y}(t).$$

Учитывая равенство $\frac{d\mathbf{z}}{dt} = -\frac{d\mathbf{y}}{dt}$, из совместного рассмотрения (3.1) и (3.25) выясняем, что $\mathbf{z}(t)$ удовлетворяет однородному дифференциальному уравнению

$$\frac{d\mathbf{z}}{dt} = -\mathbf{A}\mathbf{z}(t)$$

с начальным условием $\mathbf{z}(0) = \mathbf{x} - \mathbf{x}^{(0)}$. Решением этой начальной задачи служит вектор

$$\mathbf{z}(t) = e^{-\mathbf{A}t} \cdot \mathbf{z}(0),$$

и если спектр $\mathbf{A}$ лежит в правой полуплоскости (в частности, если, например, матрица $\mathbf{A}$ положительно определена), то $\mathbf{z}(t) \xrightarrow[t \to \infty]{} \mathbf{0}$ при любых $\mathbf{z}(0)$. Таким образом, решение $\mathbf{x}$ системы (3.1) (стационарной задачи) может быть получено как предел при $t \to \infty$ решения $\mathbf{y}(t)$ задачи Коши (3.25) (эволюционной за-

дачи) с произвольным начальным вектором $\mathbf{x}^{(0)}$.

Методы приближенного решения стационарных задач, основанные на нахождении решений нестационарных задач, асимптотически эквивалентных данным задачам для достаточно больших значений искусственной скалярной переменной, называются ***методами установления****).

Будем далее считать параметром скалярную величину τ_k, которую применительно к задаче (3.25) можно интерпретировать как шаг (вообще говоря, переменный), с которым на полуоси $[0, +\infty)$ фиксируются точки

$$t_0(=0),\ t_1,\ t_2,\ \ldots,$$

т.е. $t_{k+1} = t_k + \tau_k$, где $k = 0, 1, 2, \ldots$.

При «замораживании» $t = t_k$ уравнение (3.25) принимает вид

$$\left.\frac{d\mathbf{y}}{dt}\right|_{t=t_k} = -\mathbf{A}\mathbf{y}(t_k) + \mathbf{b}. \tag{3.26}$$

Для производной в его левой части при малых τ_k на основе определения можно записать приближенное равенство

$$\left.\frac{d\mathbf{y}}{dt}\right|_{t=t_k} = \lim_{\tau_k \to 0} \frac{\mathbf{y}(t_k + \tau_k) - \mathbf{y}(t_k)}{\tau_k} \approx \frac{\mathbf{y}(t_{k+1}) - \mathbf{y}(t_k)}{\tau_k}.$$

Теперь ясно, что полагая $\mathbf{x}^{(k)} := \mathbf{y}(t_k)$ (заметим, что $\mathbf{y}(t_0) = \mathbf{y}(0) = \mathbf{x}^{(0)}$), равенство (3.26) можно приближенно заменить равенством

$$\frac{\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}}{\tau_k} = -\mathbf{A}\mathbf{x}^{(k)} + \mathbf{b}, \tag{3.27}$$

которое можно рассматривать как некий ***явный*** итерационный процесс. Его называют ***двухслойным****) ***итерационным методом*** [161] или ***методом Ричардсона*** [158].

Более общий вид семейство двухслойных итерационных методов примет, если ввести в (3.27) невырожденный матричный параметр $\mathbf{B}_k$:

$$\mathbf{B}_k \frac{\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}}{\tau_k} = -\mathbf{A}\mathbf{x}^{(k)} + \mathbf{b}. \tag{3.28}$$

Различные конкретные итерационные процессы решения СЛАУ (3.1) (в том числе и все рассмотренные выше) получаются из (3.28) фиксированием матриц $\mathbf{B}_k$ и скаляров τ_k. При этом, если $\mathbf{B}_k$ и τ_k не зависят от k, т.е. одни и те же на каждой итерации, то (3.28) определяет ***стационарный метод***, в противном случае — ***нестационарный***. В общем случае, за исключением $\mathbf{B}_k \equiv \mathbf{E}$, (3.28) — ***неявный метод***.

Выбор параметров $\mathbf{B}_k$, τ_k в (3.28) осуществляют, добиваясь удовлетворения каких-либо отдельных или совокупности нескольких, возможно в чем-то противоречивых требований таких, как простота, хорошая структура и легкая обращаемость матриц $\mathbf{B}_k$, и в то же время, как можно более быстрая сходимость последовательности $(\mathbf{x}^{(k)})$ к решению $\mathbf{x}^*$ системы (3.1). Разумеется, оптимальность или, скорее, квазиоптимальность некоторых методов рассматриваемого семейства удается установить лишь при очень жестких ограничениях на решаемую систему (3.1).

Так, например, доказано [158, 161], что если система (3.1) — нормальная с известными границами $\lambda_{\min} > 0$, $\lambda_{\max} > 0$ спектра ее матрицы коэффициентов, то при заранее зафиксированном (максимальном в реализуемом процессе) числе итераций K метод (3.27) будет обеспечивать наименьшую погрешность, иначе, минимизировать величину $\left\| \mathbf{x}^* - \mathbf{x}^{(K)} \right\|$, в том случае, когда параметры τ_k вычисляются по формуле

$$\tau_k = \frac{2}{(\lambda_{\max} + \lambda_{\min}) + (\lambda_{\max} - \lambda_{\min}) t_{k+1}}, \tag{3.29}$$

---

*) Иногда к дифференциальным уравнениям переходят не от исходной стационарной задачи, а от какого-то конкретного итерационного метода ее решения. Получающуюся при этом асимптотически эквивалентную дифференциальную задачу называют ***непрерывным аналогом*** соответствующего итерационного метода (см., например, [36, 48]).

*) Смысл термина «двухслойный» становится понятным при изучении численных процессов решения уравнений математической физики (см. гл.20). Изучая же численное интегрирование систем дифференциальных уравнений (§ 15.5), обнаруживаем, что (3.27) есть не что иное, как явный метод Эйлера (с переменным шагом) для задачи (3.25).

где $k = 0, 1, \ldots, K-1$, а $t_k = \cos\frac{(2k-1)\pi}{2K}$ — корни полинома Чебышева K-й степени (см. § 9.1). Совокупность формул (3.27), (3.29) называют ***явным итерационным методом с чебышевским набором параметров***. Имеется обобщение приведенного утверждения и на неявный случай.

Дальнейшее формальное развитие методы установления получают как сугубо неявные методы вида (3.28) с матрицами $\mathbf{B}_k$, представляемыми в виде произведения простых легко обращаемых (например, ленточных) матриц, в связи с чем такие методы называются ***методами расщепления***. Из методов расщепления наиболее известными являются ***методы переменных направлений****) и ***попеременно-треугольный метод***. Неформальное изучение этих методов более целесообразно по месту их применения: при численном решении многомерных задач математической физики (см. далее гл.21).

Рассматриваются также ***трехслойные итерационные методы*** (в частности, с чебышевскими параметрами), связывающие уже не два, а три соседних приближения: $\mathbf{x}^{(k+1)}$, $\mathbf{x}^{(k)}$ и $\mathbf{x}^{(k-1)}$. В отличие от предыдущих, такие методы являются ***двухшаговыми***.

Другой большой класс методов итерационного решения СЛАУ (3.1) — это так называемые ***методы вариационного типа***. К ним относятся методы минимальных невязок, минимальных поправок, минимальных итераций, наискорейшего спуска, сопряженных градиентов и т.п. Хорошего понимания и обоснования таких методов можно достигнуть лишь с привлечением теории оптимизации, ибо решение линейной алгебраической системы здесь подменяется решением эквивалентной экстремальной задачи.

А именно, пусть $\mathbf{Ax} = \mathbf{b}$ — нормальная n-мерная система, т.е. $\mathbf{A}$ — положительно определенная симметричная матрица, и пусть $(\cdot,\cdot)$ — скалярное произведение в пространстве $\mathbf{R}_n$. Образуем квадратичный функционал

$$\Phi(\mathbf{x}) = (\mathbf{Ax}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x}) + c, \tag{3.30}$$

где $c \in \mathbf{R}_1$ — произвольная постоянная. Задача решения нормальной системы (3.1) и задача минимизации функционала (3.30) эквивалентны ([152 и др.]). Действительно, нормальная система

*) В зарубежной литературе используется аббревиатура ADI — *Alternating Direction Implicite* [43, 137].

имеет и притом единственное решение; обозначим его $\mathbf{x}^*$. Тогда при любом векторе $\mathbf{x} = \mathbf{x}^* + \mathbf{\Delta}$

$$\begin{aligned}\Phi(\mathbf{x}) &= \Phi(\mathbf{x}^* + \mathbf{\Delta}) = (\mathbf{A}(\mathbf{x}^* + \mathbf{\Delta}),\ \mathbf{x}^* + \mathbf{\Delta}) - 2(\mathbf{b},\ \mathbf{x}^* + \mathbf{\Delta}) + c = \\ &= (\mathbf{Ax}^*,\ \mathbf{x}^*) + (\mathbf{A\Delta},\ \mathbf{x}^*) + (\mathbf{Ax}^*,\ \mathbf{\Delta}) + (\mathbf{A\Delta},\ \mathbf{\Delta}) - 2(\mathbf{b},\ \mathbf{x}^*) - 2(\mathbf{b},\ \mathbf{\Delta}) + c = \\ &= \Phi(\mathbf{x}^*) + (\mathbf{A\Delta},\ \mathbf{x}^*) + (\mathbf{Ax}^*,\ \mathbf{\Delta}) - 2(\mathbf{Ax}^*,\ \mathbf{\Delta}) + (\mathbf{A\Delta},\ \mathbf{\Delta}) = \\ &= \Phi(\mathbf{x}^*) + (\mathbf{A\Delta},\ \mathbf{\Delta}) > \Phi(\mathbf{x}^*),\end{aligned}$$

в силу самосопряженности и положительности $\mathbf{A}$; значит,

$$\Phi(\mathbf{x}^*) = \min_{\mathbf{x} \in R_n} \Phi(\mathbf{x}).$$

Теперь можно применять различные методы численной минимизации функционала $\Phi(\mathbf{x})$ (в конечном итоге, функции n переменных $x_1, x_2, \ldots, x_n$).

Одним из наиболее популярных и хорошо разработанных методов подобного типа является ***метод сопряженных градиентов***. Приведем без вывода алгоритм, быть может, недостаточно подробный, но вполне определенный, чтобы с его помощью можно было решать нормальные СЛАУ (3.1) таким способом [138]. Фигурирующим в нем переменным можно придать следующий оптимизационный смысл:

$\mathbf{x}^{(k)}$ — k-е приближение к искомому решению $\mathbf{x}^*$;

$\xi^{(k)}$ — невязка k-го приближения, играющая роль антиградиента функции $\Phi(\mathbf{x})$;

$\mathbf{p}^{(k)}$ — направление минимизации функции $\Phi(\mathbf{x})$ в точке $\mathbf{x}^{(k)}$;

α_k — коэффициент, обеспечивающий минимум $\Phi(\mathbf{x})$ в направлении вектора $\mathbf{p}^{(k)}$;

$-\beta_k$ — коэффициент при $\mathbf{p}^{(k)}$ в формуле для вычисления направления $\mathbf{p}^{(k+1)}$, обеспечивающий A-сопряженность векторов $\mathbf{p}^{(k)}$ и $\mathbf{p}^{(k+1)}$ ($\mathbf{q}^{(k)}$ — вспомогательный вектор).

Алгоритм МСГ

Шаг 1.1. Задать $\mathbf{x}^{(0)}$ (начальный вектор) и число $\varepsilon > 0$ (допустимый уровень абсолютных погрешностей).

Шаг 1.2. Вычислить вектор $\xi^{(0)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}$ (невязка начального приближения).

Шаг 1.3. Положить $\mathbf{p}^{(0)} = \xi^{(0)}$, $k = 0$ (номер итерации).

Шаг 2.1. Вычислить вектор $\mathbf{q}^{(k)} = \mathbf{A}\mathbf{p}^{(k)}$.

Шаг 2.2. Вычислить скаляр $\alpha_k = \left(\xi^{(k)},\ \mathbf{p}^{(k)}\right) / \left(\mathbf{q}^{(k)},\ \mathbf{p}^{(k)}\right)$ (шаговый множитель).

Шаг 2.3. Вычислить вектор $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{p}^{(k)}$ (очередное приближение).

Шаг 2.4. Вычислить вектор $\xi^{(k+1)} = \xi^{(k)} - \alpha_k \mathbf{q}^{(k)}$ (невязка $(k+1)$-го приближения*)).

Шаг 2.5. Проверить выполнение неравенства $\left\|\xi^{(k+1)}\right\|_2 \le \varepsilon$; если «да», остановить работу алгоритма и вывести результаты.

Шаг 3.1. Вычислить скаляр $\beta_k = \left(\xi^{(k+1)}, \mathbf{q}^{(k)}\right) / \left(\mathbf{p}^{(k)}, \mathbf{q}^{(k)}\right)$.

Шаг 3.2. Вычислить вектор $\mathbf{p}^{(k+1)} = \xi^{(k+1)} - \beta_k \mathbf{p}^{(k)}$ (новое направление минимизации).

Шаг 3.3. Положить $k := k + 1$ и вернуться к шагу 2.1.

Интересно определить место, которое занимает этот метод в общей классификации методов решения линейных алгебраических систем. Дело в том, что метод сопряженных градиентов, являясь по форме итерационным, фактически должен быть отнесен к прямым методам, ибо доказано, что с его помощью минимум квадратичной функции (3.30) от n переменных, иначе, решение n-мерной линейной системы (3.1), достигается ровно за n шагов при любом начальном векторе $\mathbf{x}^{(0)}$. Применяют же метод сопряженных градиентов именно как итерационный метод (что видно и из приведенного алгоритма), имея в виду два обстоятельства. Во-первых, реальный вычислительный процесс может

*) Полагая $\xi^{(k)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}$, видим, что, в силу 2.3 и 2.1,

$$\xi^{(k+1)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k+1)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)} - \mathbf{A}\alpha_k \mathbf{p}^{(k)} = \xi^{(k)} - \alpha_k \mathbf{A}\mathbf{p}^{(k)} = \xi^{(k)} - \alpha_k \mathbf{q}^{(k)}.$$

Использование такого выражения невязки $\xi^{(k+1)}$ позволяет обходиться без вычисления вектора $\mathbf{A}\mathbf{x}^{(k+1)}$.Однако нужно понимать, что подобная экономия в арифметических операциях может отразиться на вычислительной устойчивости метода.

Во-первых, реальный вычислительный процесс может быть довольно далек от идеального и, вследствие неизбежных ошибок округления, на n-м шаге может быть не достигнута нужная точность. Во-вторых, если размерность n решаемой задачи велика, то число шагов, достаточное для получения решения системы с нужной точностью (т.е. выход по критерию 2.4), может оказаться значительно меньшим этой (n) теоретической величины.

Покажем сначала возможности метода сопряженных градиентов на очень простом примере, где точное решение найдется раньше, чем будет выполнен полный цикл предложенного алгоритма.

Пример 3.4. Для решения системы

$$\begin{cases} 2x_1 + x_2 = 1, \\ x_1 + 2x = -1 \end{cases}$$

с положительно определенной матрицей $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ и правой частью $\mathbf{b} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ возьмем начальное приближение $\mathbf{x}^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. Его **невязка** $\xi^{(0)} := \mathbf{b} - \mathbf{A}\,\mathbf{x}^{(0)} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. В соответствии с алгоритмом далее имеем:

$$\mathbf{p}^{(0)} := \xi^{(0)} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad \mathbf{q}^{(0)} := \mathbf{A}\,\mathbf{p}^{(0)} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

и, следовательно,

$$\alpha_0 := \frac{\left(\xi^{(0)},\ \mathbf{p}^{(0)}\right)}{\left(\mathbf{q}^{(0)},\ \mathbf{p}^{(0)}\right)} = 1.$$

Таким образом, находим первое приближение

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \alpha_0 \mathbf{p}^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + 1 \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

с невязкой

$$\xi^{(1)} = \xi^{(0)} - \alpha_0 \mathbf{q}^{(0)} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} - \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

говорящей о том, что $\mathbf{x}^* = \mathbf{x}^{(1)} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

Теперь рассмотрим пример более типичного поведения метода сопряженных градиентов, применяя его к трехмерной СЛАУ с симметричной положительно определенной матрицей коэффициентов.

Пример 3.5. Дана система $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, где

$$\mathbf{A} = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & -1 \\ 1 & -1 & 2 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}.$$

Продемонстрируем все расчеты, выполняя шаг за шагом предписания алгоритма сопряженных градиентов, игнорируя лишь шаг 2.5, поскольку задавать ε здесь нет смысла ввиду малой размерности системы.

Приняв за начальное приближение $\mathbf{x}^{(0)}$ нуль-вектор, далее последовательно вычисляем:

$$\xi^{(0)} := \mathbf{b} - \mathbf{A}\,\mathbf{x}^{(0)} = \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}, \quad \mathbf{p}^{(0)} := \xi^{(0)} = \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}, \quad \mathbf{q}^{(0)} := \mathbf{A}\,\mathbf{p}^{(0)} = \begin{pmatrix} 0 \\ 4 \\ -5 \end{pmatrix},$$

$$\alpha_0 := \frac{\left(\xi^{(0)},\ \mathbf{p}^{(0)}\right)}{\left(\mathbf{q}^{(0)},\ \mathbf{p}^{(0)}\right)} = \frac{9}{18} = 0.5, \quad \mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \alpha_0 \mathbf{p}^{(0)} = \begin{pmatrix} 0.5 \\ 1 \\ -1 \end{pmatrix},$$

$$\xi^{(1)} = \xi^{(0)} - \alpha_0 \mathbf{q}^{(0)} = \begin{pmatrix} 1 \\ 0 \\ 0.5 \end{pmatrix}, \quad \beta_0 := \frac{\left(\xi^{(1)},\ \mathbf{q}^{(0)}\right)}{\left(\mathbf{q}^{(0)},\ \mathbf{p}^{(0)}\right)} = -\frac{5}{36},$$

$$\mathbf{p}^{(1)} = \xi^{(1)} - \beta_0\,\mathbf{p}^{(0)} = \frac{1}{36}\cdot\begin{pmatrix} 41 \\ 10 \\ 8 \end{pmatrix}, \quad \mathbf{q}^{(1)} = \mathbf{A}\,\mathbf{p}^{(1)} = \frac{1}{36}\cdot\begin{pmatrix} 90 \\ 2 \\ 47 \end{pmatrix},$$

$$\alpha_1 = \frac{\left(\xi^{(1)},\ \mathbf{p}^{(1)}\right)}{\left(\mathbf{q}^{(1)},\ \mathbf{p}^{(1)}\right)} = \frac{90}{227} \approx 0.396476, \quad \mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \alpha_1 \mathbf{p}^{(1)} = \begin{pmatrix} 0.951542 \\ 1.110132 \\ -0.911894 \end{pmatrix},$$

$$\xi^{(2)} = \xi^{(1)} + \alpha_1 \mathbf{q}^{(1)} \approx \begin{pmatrix} 0.008810 \\ -0.022026 \\ -0.017621 \end{pmatrix}, \quad \beta_1 = \frac{\left(\xi^{(2)},\ \mathbf{q}^{(1)}\right)}{\left(\mathbf{q}^{(1)},\ \mathbf{p}^{(1)}\right)} \approx -0.000699,$$

$$\mathbf{p}^{(2)} = \xi^{(2)} - \beta_1 \mathbf{p}^{(1)} \approx \begin{pmatrix} 0.009606 \\ -0.021832 \\ -0.017466 \end{pmatrix}, \quad \mathbf{q}^{(2)} = \mathbf{A}\,\mathbf{p}^{(2)} \approx \begin{pmatrix} 0.001746 \\ -0.004366 \\ -0.003494 \end{pmatrix},$$

$$\alpha_2 = \frac{\left(\xi^{(2)},\ \mathbf{p}^{(2)}\right)}{\left(\mathbf{q}^{(2)},\ \mathbf{p}^{(2)}\right)} \approx 5.044386, \quad \mathbf{x}^{(3)} = \mathbf{x}^{(2)} + \alpha_2 \mathbf{p}^{(2)} = \begin{pmatrix} 0.999998 \\ 1.000003 \\ -0.999999 \end{pmatrix}.$$

Отличие вектора $\mathbf{x}^{(3)}$ от истинного решения $\mathbf{x}^* = (1;\ 1;\ -1)^T$ обусловлено лишь ошибками округления, а о том, что $\mathbf{x}^{(3)}$ действительно можно

считать искомым решением с определенной точностью (не зная $\mathbf{x}^*$), следует судить, выполнив еще один шаг, а именно, подсчитав невязку $\xi^{(3)}$.

Простейший вариант ***метода минимальных невязок*** определяется совокупностью формул

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \tau_k \xi^{(k)}, \quad \xi^{(k)} = \mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}, \quad \tau_k = \frac{\left(\mathbf{A}\xi^{(k)},\ \xi^{(k)}\right)}{\left(\mathbf{A}\xi^{(k)},\ \mathbf{A}\xi^{(k)}\right)}.$$

Его можно рассматривать как явный двухслойный итерационный процесс (3.27), в котором параметр τ_k на каждом итерационном шаге $k = 0, 1, 2, \ldots$ выбирается таким, чтобы минимизировалась евклидова норма невязки $\xi^{(k+1)}$ получаемого приближения $\mathbf{x}^{(k+1)}$.

Действительно, вычтем из вектора $\xi^{(k+1)} = \mathbf{A}\mathbf{x}^{(k+1)} - \mathbf{b}$ вектор $\xi^{(k)}$. Имеем

$$\xi^{(k+1)} - \xi^{(k)} = \mathbf{A}\mathbf{x}^{(k+1)} - \mathbf{A}\mathbf{x}^{(k)} = \mathbf{A}\mathbf{x}^{(k)} - \mathbf{A}\tau_k \xi^{(k)} - \mathbf{A}\mathbf{x}^{(k)},$$

т.е.

$$\xi^{(k+1)} = \xi^{(k)} - \tau_k \mathbf{A}\xi^{(k)}.$$

Возводя последнее равенство в квадрат (в смысле скалярного умножения векторов), получаем

$$\left(\xi^{(k+1)},\ \xi^{(k+1)}\right) = \left(\xi^{(k)},\ \xi^{(k)}\right) - 2\tau_k\left(\mathbf{A}\xi^{(k)},\ \xi^{(k)}\right) + \tau_k^2\left(\mathbf{A}\xi^{(k)},\ \mathbf{A}\xi^{(k)}\right)$$

или, что то же,

$$\left\|\xi^{(k+1)}\right\|^2 = \left\|\xi^{(k)}\right\|^2 - 2\tau_k\left(\mathbf{A}\xi^{(k)},\ \xi^{(k)}\right) + \tau_k^2 \cdot \left\|\mathbf{A}\xi^{(k)}\right\|^2.$$

Легко видеть, что минимум этой положительной квадратичной функции (значит, и величины $\left\|\xi^{(k+1)}\right\|$) достигается именно при указанном в записи метода значении τ_k.

В случае нормальной системы для метода минимальных невязок можно получить ту же оценку скорости сходимости, что и для метода простой итерации

$$\mathbf{x}^{(k+1)} = (\mathbf{E} - \tau\mathbf{A})\mathbf{x}^{(k)} + \tau\mathbf{b}$$

при оптимальном значении параметра $\tau = \dfrac{2}{\lambda_{\min} + \lambda_{\max}}$ (в предположении, что известны границы $\lambda_{\min}$ и $\lambda_{\max}$ спектра матрицы $\mathbf{A}$) [158].

Рассмотренные здесь методы далеко не исчерпывают все многообразие итерационных способов решения СЛАУ. В частности, нами совсем не затрагивалась проблема решения больших разреженных систем, где на первый план выходят блочные методы, максимально сохраняющие исходную разреженность матриц (см., например, [29, 64, 197]).

3.6. БЫСТРОСХОДЯЩИЙСЯ ИТЕРАЦИОННЫЙ СПОСОБ ОБРАЩЕНИЯ МАТРИЦ

Согласно леммам 3.1, 3.2 (см. § 3.1), если матрица $\mathbf{B}=\mathbf{E}-\mathbf{A}$ мала (в смысле ее нормы или собственных значений), то обратная к $\mathbf{A}$ матрица

$$\mathbf{A}^{-1}=(\mathbf{E}-\mathbf{B})^{-1}=\mathbf{E}+\mathbf{B}+\mathbf{B}^2+\ldots,$$

в принципе, может быть найдена сколь угодно точно приближенным суммированием данного матричного ряда. Однако такой непосредственный подход к вычислению имеет два очевидных недостатка: во-первых, реально его можно применить лишь для обращения матриц, близких к единичной, во-вторых, сходимость последовательностей частичных сумм этого ряда будет медленной даже при достаточно малых нормах матриц $\mathbf{B}$. Поэтому, пользуясь отмеченным фактом лишь как теоретической основой, построим итерационный процесс, определяющий существенно более быстро сходящуюся последовательность приближений к обратной для $\mathbf{A}$ матрице $\mathbf{A}^{-1}$. Будем далее обозначать эти приближения, получаемые на k-м шаге, через $\mathbf{U}_k$, а их ***невязки*** $\mathbf{E}-\mathbf{A}\mathbf{U}_k$ — через $\mathbf{\Psi}_k$.

Лемма 3.3. *Если для матрицы* $\mathbf{A}$ *найдется такая обратимая матрица* $\mathbf{U}_0$, *что модули всех собственных чисел матрицы* $\mathbf{\Psi}_0=\mathbf{E}-\mathbf{A}\mathbf{U}_0$ *меньше единицы, то матрица* $\mathbf{A}$ *обратима и для обратной матрицы справедливо представление*

$$\mathbf{A}^{-1}=\mathbf{U}_0(\mathbf{E}-\mathbf{\Psi}_0)^{-1}=\mathbf{U}_0\left(\mathbf{E}+\mathbf{\Psi}_0+\mathbf{\Psi}_0^2+\ldots\right). \quad (3.31)$$

Доказательство. Из равенства

$$\mathbf{A}\mathbf{U}_0=\mathbf{E}-\mathbf{\Psi}_0, \quad (3.32)$$

в силу обратимости матриц $\mathbf{U}_0$ и $\mathbf{E}-\mathbf{\Psi}_0$ (последнее по

лемме 3.1), имеем

$$\mathbf{A}=(\mathbf{E}-\mathbf{\Psi}_0)\mathbf{U}_0^{-1}=\left(\left((\mathbf{E}-\mathbf{\Psi}_0)^{-1}\right)^{-1}\mathbf{U}_0^{-1}\right)=\left(\mathbf{U}_0(\mathbf{E}-\mathbf{\Psi}_0)^{-1}\right)^{-1},$$

т.е. матрица $\mathbf{A}$ обратима и справедливо представление

$$\mathbf{A}^{-1}=\mathbf{U}_0(\mathbf{E}-\mathbf{\Psi}_0)^{-1}.$$

Доказательство завершается разложением $(\mathbf{E}-\mathbf{\Psi}_0)^{-1}$ в матричный ряд (лемма 3.1).

Очевидным следствием лемм 3.2 и 3.3 является следующая лемма.

Лемма 3.4. *Пусть матрица* $\mathbf{U}_0$ *обратима и* $\|\mathbf{\Psi}_0\|<1$. *Тогда*:

1) *существует матрица* $\mathbf{A}^{-1}$;
2) *справедливо представление* $\mathbf{A}^{-1}$ *по формуле* (3.31);
3) *имеет место оценка* $\left\|\mathbf{A}^{-1}\right\|\le\dfrac{\|\mathbf{U}_0\|}{1-\|\mathbf{\Psi}_0\|}$.

Для построения итерационного процесса зафиксируем в разложении (3.31) $m+1$ первых слагаемых и будем считать первым приближением к $\mathbf{A}^{-1}$ матрицу

$$\mathbf{U}_1=\mathbf{U}_0\left(\mathbf{E}+\mathbf{\Psi}_0+\ldots+\mathbf{\Psi}_0^m\right).$$

Найдем выражение невязки $\mathbf{\Psi}_1$ этого приближения через невязку $\mathbf{\Psi}_0$ предыдущего (в данном случае начального) приближения $\mathbf{U}_0$:

$$\begin{aligned}\mathbf{\Psi}_1&=\mathbf{E}-\mathbf{A}\mathbf{U}_1=\mathbf{E}-\mathbf{A}\mathbf{U}_0\left(\mathbf{E}+\mathbf{\Psi}_0+\ldots\mathbf{\Psi}_0^m\right)=\\&=\mathbf{E}-(\mathbf{E}-\mathbf{\Psi}_0)\left(\mathbf{E}+\mathbf{\Psi}_0+\ldots+\mathbf{\Psi}_0^m\right)=\mathbf{E}-\left(\mathbf{E}-\mathbf{\Psi}_0^{m+1}\right)=\mathbf{\Psi}_0^{m+1}.\end{aligned} \quad (3.33)$$

Благодаря полученной связи между невязками, можно утверждать, что если выполняются условия лемм 3.3 или 3.4 по отношению к матрицам $\mathbf{U}_0$, $\mathbf{\Psi}_0$, то для матриц $\mathbf{U}_1$, $\mathbf{\Psi}_1$ они тем более будут выполнены. Следовательно, к матрицам $\mathbf{U}_1$, $\mathbf{\Psi}_1$, можно применить все рассуждения, проведенные выше для $\mathbf{U}_0$, $\mathbf{\Psi}_0$. Таким образом, приходим к итерационному процессу

$$\begin{cases}\mathbf{\Psi}_k=\mathbf{E}-\mathbf{A}\mathbf{U}_k,\\ \mathbf{U}_{k+1}=\mathbf{U}_k\left(\mathbf{E}+\mathbf{\Psi}_k+\ldots+\mathbf{\Psi}_k^m\right),\end{cases} \quad (3.34)$$

где $k=0,1,2,\ldots$ — номер итерации; $\mathbf{U}_0$ — задаваемая начальная

матрица, близкая к $\mathbf{A}^{-1}$ в указанном выше смысле, а $m \in \mathrm{N}$ — параметр метода.

Изучим сходимость этого процесса.

Теорема 3.13. *Пусть квадратные матрицы* $\mathbf{A}$ *и* $\mathbf{U}_0$ *таковы, что матрица* $\mathbf{U}_0$ *обратима и* $\|\mathbf{\Psi}_0\| < 1$. *Тогда существует обратная к* $\mathbf{A}$ *матрица* $\mathbf{A}^{-1}$ *и к ней сходится последовательность матриц* $\mathbf{U}_k$, *определяемая итерационным процессом* (3.34). *При этом имеет место точное равенство*

$$\mathbf{A}^{-1} - \mathbf{U}_k = \left(\mathbf{A}^{-1} - \mathbf{U}_0\right)\mathbf{\Psi}_0^{(m+1)^k - 1} \tag{3.35}$$

и справедливы оценки погрешности:

1) $\left\|\mathbf{A}^{-1} - \mathbf{U}_k\right\| \le \dfrac{\|\mathbf{U}_k \mathbf{\Psi}_k\|}{1 - \|\mathbf{\Psi}_k\|}$;

2) $\left\|\mathbf{A}^{-1} - \mathbf{U}_k\right\| \le \dfrac{\|\mathbf{U}_0\|}{1 - \|\mathbf{\Psi}_0\|} \cdot \|\mathbf{\Psi}_0\|^{(m+1)^k}$.

Доказательство. Существование $\mathbf{A}^{-1}$ следует из леммы 3.4. Упомянутая повторяемость рассуждений и выкладок, проведенных на первом итерационном шаге, позволяет считать очевидными равенства типа (3.31), (3.33) для k-й итерации:

$$\mathbf{A}^{-1} = \mathbf{U}_k(\mathbf{E} - \mathbf{\Psi}_k)^{-1} = \mathbf{U}_k\left(\mathbf{E} + \mathbf{\Psi}_k + \mathbf{\Psi}_k^2 + \ldots\right), \tag{3.36}$$

$$\mathbf{\Psi}_k = \mathbf{E} - \mathbf{A}\mathbf{U}_k = \mathbf{\Psi}_{k-1}^{m+1} = \mathbf{\Psi}_{k-2}^{(m+1)^2} = \ldots = \mathbf{\Psi}_0^{(m+1)^k}. \tag{3.37}$$

Из (3.31) имеем

$$\begin{aligned}\mathbf{A}^{-1} - \mathbf{U}_0 &= \mathbf{U}_0\left(\mathbf{E} + \mathbf{\Psi}_0 + \mathbf{\Psi}_0^2 + \ldots\right) - \mathbf{U}_0 = \\ &= \mathbf{U}_0\left(\mathbf{E} + \mathbf{\Psi}_0 + \mathbf{\Psi}_0^2 + \ldots\right)\mathbf{\Psi}_0 = \mathbf{A}^{-1}\mathbf{\Psi}_0,\end{aligned} \tag{3.38}$$

а из (3.36) аналогично (с учетом (3.37)) получаем

$$\mathbf{A}^{-1} - \mathbf{U}_k = \mathbf{A}^{-1}\mathbf{\Psi}_k = \mathbf{A}^{-1}\mathbf{\Psi}_0^{(m+1)^k}. \tag{3.39}$$

Заменяя здесь в правой части $\mathbf{A}^{-1}\mathbf{\Psi}_0$ на $\mathbf{A}^{-1} - \mathbf{U}_0$ (см. (3.38)), получаем утверждаемое в теореме равенство (3.35). Переходя в нем к нормам, в соответствии с условием заключаем, что

$$\left\|\mathbf{A}^{-1} - \mathbf{U}_k\right\| \le \left\|\mathbf{A}^{-1} - \mathbf{U}_0\right\| \cdot \|\mathbf{\Psi}_0\|^{(m+1)^k - 1} \xrightarrow[k \to \infty]{} 0,$$

т.е. имеет место сходимость последовательности $(\mathbf{U}_k)_{k=1}^{\infty}$ к матрице $\mathbf{A}^{-1}$ по норме, а значит, и поэлементная сходимость.

Для доказательства первой оценки (апостериорной) вычтем $\mathbf{U}_k$ из (3.36):

$$\mathbf{A}^{-1} - \mathbf{U}_k = \mathbf{U}_k\left(\mathbf{E} + \mathbf{\Psi}_k + \mathbf{\Psi}_k^2 + \ldots\right) - \mathbf{U}_k = \mathbf{U}_k\mathbf{\Psi}_k(\mathbf{E} - \mathbf{\Psi}_k)^{-1}.$$

Отсюда по лемме 3.2 с учетом (3.37) получаем требуемую оценку 1).

Вторая оценка (априорная) может быть найдена в результате загрубления первой. Но можно вывести ее непосредственно из равенства (3.39), подставив в его правую часть вместо $\mathbf{A}^{-1}$ выражение $\mathbf{U}_0(\mathbf{E} - \mathbf{\Psi}_0)^{-1}$ (см. (3.31)):

$$\mathbf{A}^{-1} - \mathbf{U}_k = \mathbf{U}_0(\mathbf{E} - \mathbf{\Psi}_0)^{-1}\mathbf{\Psi}_0^{(m+1)^k}.$$

Переход к нормам в последнем равенстве и привлечение леммы 3.2 завершает доказательство теоремы.

Равенства (3.34) определяют фактически не один, а целое семейство итерационных методов обращения. Фиксированием параметра $m = 1, 2, \ldots$ можно получать конкретные процессы $(m+1)$-го порядка скорости сходимости*). Этот порядок может быть сколь угодно большим, однако обычно ограничиваются процессами второго $(m = 1)$ и третьего $(m = 2)$ порядков. Приоритет процесса второго порядка связан с его простотой и более ранним появлением: первая публикация об этом методе относится к 1933 г. и принадлежит Г. Шульцу [205], в связи с чем и все семейство (3.34) естественно называть ***методом Шульца****). Метод третьего порядка целесообразно использовать из тех соображений, что он, как показал М. Альтман [201], обладает свойством минимальности вычислительных затрат, требующихся для обращения матриц с заданной точностью методами семейства (3.34).

Отметим, что как сам быстросходящийся итерационный процесс (3.34), так и представленные теоремой 3.11 результаты можно без каких-либо особых дополнительных условий отнести к более общей задаче обращения линейных ограниченных операторов в полных нормированных пространствах.

---

*) Определение порядка итерационного процесса см. далее в § 5.3.

**) В разных литературных источниках можно встретить и другие названия этого метода: *Хотеллинга*, *Бодевига* (Бодвига), а также *Hobo*.

Процесс (3.34) построения приближений к обратной матрице легко видоизменить подобно тому, как это было сделано с методом простых итераций решения СЛАУ, когда для более оперативного учета получаемой на текущей итерации информации перешли от него к методу Зейделя (см. § 3.3). Например, ***зейделева модификация метода Шульца второго порядка*** может быть определена равенствами

$$\begin{cases} \mathbf{\Psi}_k = \mathbf{E} - \mathbf{A}\mathbf{U}_k, \\ \mathbf{U}_{k+1} = \mathbf{U}_k + \mathbf{U}_k \underline{\mathbf{\Psi}_k} + \mathbf{U}_{k+1} \overline{\mathbf{\Psi}_k}, \end{cases} \tag{3.40}$$

где $k = 0, 1, 2, \ldots$; $\mathbf{\Psi}_k = \underline{\mathbf{\Psi}_k} + \overline{\mathbf{\Psi}_k}$, а $\underline{\mathbf{\Psi}_k}$ и $\overline{\mathbf{\Psi}_k}$ — соответственно нижняя треугольная и строго верхняя треугольная матрицы [36]. При реализации этой модификации нужно либо расписывать формулы (3.40) поэлементно (чтобы не работать с заведомо нулевыми элементами), либо формировать матрицу $\mathbf{U}_{k+1}$ постепенным замещением старых элементов новыми, осуществляя на k-й итерации цикл присвоений

$$\mathbf{U} := \mathbf{U} + \mathbf{U}\mathbf{\Psi},$$

где до начала цикла в правой части в двумерном массиве $\mathbf{U}$ должна содержаться матрица $\mathbf{U}_k$, а в двумерном массиве $\mathbf{\Psi}$ — матрица $\mathbf{\Psi}_k$ (заполнение массивов новыми элементами производится по строкам). Процесс (3.40) при том же шаговом объеме вычислений и такой же простоте, что и в методе Шульца второго порядка, может дать определенный выигрыш в скорости сходимости. Это можно показать сравнением невязок первых приближений при тех или иных предположениях относительно начальной невязки, иначе, при тех или иных требованиях к начальной матрице $\mathbf{U}_0$ [36].

Вообще, проблема выбора начального приближения $\mathbf{U}_0$ в рассматриваемых здесь процессах итерационного обращения матриц не позволяет относиться к ним как к самостоятельным универсальным методам, конкурирующим с прямыми методами обращения, основанными, например, на LU-разложении матриц. Имеются некоторые рекомендации по выбору $\mathbf{U}_0$ (см. [20, 201] и др.), обеспечивающие выполнение условия*) $\rho(\mathbf{\Psi}_0) < 1$, являющегося необходимым и достаточным для сходимости процесса

*) Через $\rho(\cdot)$ здесь обозначается спектральный радиус указанной в скобках матрицы.

(3.34). Однако при этом, во-первых, требуется знать оценку сверху спектра обращаемой матрицы $\mathbf{A}$ либо матрицы $\mathbf{A}\mathbf{A}^T$ (а именно, если $\mathbf{A}$ — симметричная положительно определенная и $\rho(\mathbf{A}) \le \beta$, то можно взять $\mathbf{U}_0 = \alpha\mathbf{E}$, где $\alpha \in \left(0, \frac{2}{\beta}\right)$; если же $\mathbf{A}$ — произвольная невырожденная матрица и $\rho\left(\mathbf{A}\mathbf{A}^T\right) \le \beta$, то полагают $\mathbf{U}_0 = \alpha\mathbf{A}^T$, где также $\alpha \in \left(0, \frac{2}{\beta}\right)$; можно, конечно, упростить ситуацию и, воспользовавшись тем, что $\rho\left(\mathbf{A}\mathbf{A}^T\right) \le \left\|\mathbf{A}\mathbf{A}^T\right\|$, положить $\mathbf{U}_0 = \dfrac{\mathbf{A}^T}{\left\|\mathbf{A}\mathbf{A}^T\right\|}$). Во-вторых, при таком задании начальной матрицы нет гарантии, что $\|\mathbf{\Psi}_0\|$ будет малой (возможно, даже окажется $\|\mathbf{\Psi}_0\| > 1$), и высокий порядок скорости сходимости обнаружится далеко не сразу.

Все сказанное выше не означает, что подобные методы обращения матриц (и операторов) не имеют своей сферы применения. В частности, ниже (в § 7.2) будет рассматриваться способ решения систем нелинейных уравнений, базирующийся на методе Ньютона с приближенным обращением матриц Якоби по методу Шульца, а в § 18.4 метод Шульца используется как составная часть квадратурно-итерационного метода вычисления резольвент линейных интегральных уравнений.

3.7. О РОЛИ ОШИБОК ОКРУГЛЕНИЯ В ИТЕРАЦИОННЫХ МЕТОДАХ

Обратимся, наконец, к вопросам практической реализации итерационных методов решения линейных алгебраических задач.

Многие утверждения о сходимости итерационных процессов говорят о том, что решение поставленной задачи при определенных условиях может быть найдено этим процессом сколь угодно точно, причем погрешность каждого приближения может быть эффективно проконтролирована (см. теоремы 3.2, 3.6, 3.8, 3.12, а также теорему 3.3 с замечанием 3.5 и теорему 3.9 с замечанием 3.7). Нетрудно понять, что все это справедливо на самом деле лишь до тех пор, пока на погрешность метода (остаточную погрешность) не наложится вычислительная погрешность (погрешность округлений), неизбежная при любых реальных компьютерных расчетах. Особенно существенное и даже пагубное влияние на результат решения задачи итерационным методом

могут оказать ошибки округления в тех случаях, когда утверждения о сходимости метода не содержат эффективных оценок погрешности (теоремы 3.1, 3.4, 3.5, 3.10, 3.12).

Рассмотрим различие между реальным и идеальным итерационными процессами на простейшем объекте – на методе простой итерации.

Пусть на k-м итерационном шаге вычислений по методу (3.3) ошибки округлений составляют вектор $\boldsymbol{\gamma}^{(k)}$. Тогда в отличие от идеального МПИ (3.3), генерирующего последовательность приближений $\mathbf{x}^{(k)}$ к решению $\mathbf{x}^*$ системы (3.1) такому, что

$$\mathbf{x}^* = \mathbf{B}\mathbf{x}^* + \mathbf{c}, \qquad (3.41)$$

реальный МПИ будет иметь вид

$$\widetilde{\mathbf{x}}^{(k+1)} = \mathbf{B}\widetilde{\mathbf{x}}^{(k)} + \mathbf{c} + \boldsymbol{\gamma}^{(k)}. \qquad (3.42)$$

Изучим поведение векторов

$$\boldsymbol{\mu}_k := \widetilde{\mathbf{x}}^{(k)} - \mathbf{x}^*$$

— ошибок приближений $\widetilde{\mathbf{x}}^{(k)}$, получаемых реальным МПИ (3.42).

Вычитая (3.41) из (3.42), имеем

$$\widetilde{\mathbf{x}}^{(k+1)} - \mathbf{x}^* = \mathbf{B}\left(\widetilde{\mathbf{x}}^{(k)} - \mathbf{x}^*\right) + \boldsymbol{\gamma}^{(k)},$$

т.е.

$$\begin{aligned}\boldsymbol{\mu}_{k+1} &= \mathbf{B}\boldsymbol{\mu}_k + \boldsymbol{\gamma}^{(k)} = \mathbf{B}\left(\mathbf{B}\boldsymbol{\mu}_{k-1} + \boldsymbol{\gamma}^{(k-1)}\right) + \boldsymbol{\gamma}^{(k)} = \\ &= \mathbf{B}^2\left(\mathbf{B}\boldsymbol{\mu}_{k-2} + \boldsymbol{\gamma}^{(k-2)}\right) + \mathbf{B}\boldsymbol{\gamma}^{(k-1)} + \boldsymbol{\gamma}^{(k)} = \ldots = \\ &= \mathbf{B}^{k+1}\boldsymbol{\mu}_0 + \left(\mathbf{B}^k\boldsymbol{\gamma}^{(0)} + \mathbf{B}^{k-1}\boldsymbol{\gamma}^{(1)} + \ldots + \mathbf{B}\boldsymbol{\gamma}^{(k-1)} + \boldsymbol{\gamma}^{(k)}\right). \qquad (3.43)\end{aligned}$$

Первое слагаемое в последнем выражении отвечает за погрешность идеального МПИ и может быть сделано сколь угодно малым в процессе итерирования при условии $\rho(\mathbf{B}) < 1$ (см. лемму 3.1). Чтобы оценить второе слагаемое, предположим, что порог абсолютных погрешностей округлений, допускаемых на каждой итерации, есть γ, т.е.

$$\left\|\boldsymbol{\gamma}^{(k)}\right\| \le \gamma \quad \forall k \in \mathbf{N}_0.$$

Тогда

$$\left\|\mathbf{B}^k\boldsymbol{\gamma}^{(0)} + \mathbf{B}^{k-1}\boldsymbol{\gamma}^{(1)} + \ldots + \mathbf{B}\boldsymbol{\gamma}^{(k-1)} + \boldsymbol{\gamma}^{(k)}\right\| \le \gamma\left\|\mathbf{E} + \mathbf{B} + \cdots + \mathbf{B}^k\right\|,$$

и, если $\|\mathbf{B}\| \le q < 1$, то второе слагаемое в (3.43), хотя и не стремится к нулю, но ограничено по норме величиной

$$\gamma\frac{1-q^k}{1-q} < \frac{\gamma}{1-q}.$$

При условии же $\rho(\mathbf{B}) < 1$, теоретически обеспечивающем сходимость идеального МПИ (3.3), малость этого второго слагаемого отнюдь не гарантируется, что означает допустимость ситуаций, когда в ходе реальных итераций погрешность округлений будет накапливаться вплоть до переполнения множества чисел, представляемых используемым компьютером.

Более детальный анализ влияния ошибок округления на итерационный процесс с попыткой пролить свет на природу этого влияния можно найти, например, в [13]. Здесь же ограничимся напоминанием о том, что необходимо с осторожностью применять процессы, когда для них нет эффективных оценок погрешности, и по возможности, учитывать влияние ошибок округления, если такие оценки есть. Например, применительно к МПИ решения СЛАУ выше фактически доказана

Теорема 3.14.*) *Пусть $\|\mathbf{B}\| \le q < 1$ и приближения $\widetilde{\mathbf{x}}^{(k)}$ к решению $\mathbf{x}^*$ системы (3.2) получаются посредством равенства (3.42), где $\boldsymbol{\gamma}^{(k)}$ — вектор ошибок округлений таких, что $\left\|\boldsymbol{\gamma}^{(k)}\right\| \le \gamma$. Тогда погрешность k-го приближения при любом $k \in \mathbf{N}$ можно оценить неравенством*

$$\left\|\mathbf{x}^* - \widetilde{\mathbf{x}}^{(k)}\right\| \le \frac{q}{1-q}\left\|\widetilde{\mathbf{x}}^{(k)} - \widetilde{\mathbf{x}}^{(k-1)}\right\| + \frac{\gamma}{1-q}. \qquad (3.44)$$

Действительно, для последовательности $\left(\mathbf{x}^{(k)}\right)$, получаемой МПИ (3.3), справедливо равенство

$$\mathbf{x}^* - \mathbf{x}^{(k+1)} = \mathbf{B}^{k+1}\left(\mathbf{x}^* - \mathbf{x}^{(0)}\right).$$

Следовательно, считая, что процессы (3.3) и (3.42) начинаются с одного начального приближения $\mathbf{x}^{(0)} = \widetilde{\mathbf{x}}^{(0)}$, в идентичном (3.43) равенстве

$$\mathbf{x}^* - \widetilde{\mathbf{x}}^{(k+1)} = \mathbf{B}^{k+1}\left(\mathbf{x}^* - \widetilde{\mathbf{x}}^{(0)}\right) - \left(\mathbf{B}^k\boldsymbol{\gamma}^{(0)} + \mathbf{B}^{k-1}\boldsymbol{\gamma}^{(1)} + \ldots + \mathbf{B}\boldsymbol{\gamma}^{(k-1)} + \boldsymbol{\gamma}^{(k)}\right)$$

*) См. также [3].

можно заменить $\mathbf{B}^{k+1}(\mathbf{x}^* - \widetilde{\mathbf{x}}^{(0)})$ на $\mathbf{x}^* - \mathbf{x}^{(k+1)}$. Таким образом, погрешности $(k+1)$-х приближений реального (3.42) и идеального (3.3) методов различаются лишь слагаемым, оцененным выше по норме величиной $\frac{\gamma}{1-q}$, т.е. и для процесса (3.42) можно воспользоваться оценкой, выведенной в теореме 3.2.

Отметим, что как непосредственно видно из оценки (3.44) (при значениях q, приближающихся к единице), роль ошибок округлений в образовании общей погрешности тем сильнее, чем медленнее сходимость итерационного процесса.

УПРАЖНЕНИЯ

3.1. Запишите итерационный процесс Якоби нахождения решения системы

$$\begin{cases} 5x_1 + 2x_2 - x_3 + x_4 = 9, \\ x_1 - 4x_2 + 2x_4 = 10, \\ 2x_1 + 3x_2 - 9x_3 - x_4 = -10, \\ 3x_1 + x_3 - 6x_4 = -5. \end{cases}$$

Каким должен быть критерий окончания процесса итерирования, чтобы максимальная из абсолютных погрешностей компонент приближенного решения не превышала заданного малого $\varepsilon > 0$?

3.2. Проверьте, выполняются ли необходимые условия сходимости методов Якоби и Зейделя, примененных к системе

$$\begin{cases} x_1 + x_2 = 2, \\ x_1 + 2x_2 + x_3 = 4, \\ x_2 + 2x_3 = 3. \end{cases}$$

3.3. Сделайте по пять итераций методов Якоби и Зейделя для системы

$$\begin{cases} 10x_1 + x_2 - 2x_3 = 10, \\ x_1 - 5x_2 + x_3 = 10, \\ 3x_1 - x_2 + 10x_3 = -5. \end{cases}$$

Сколько верных знаков можно гарантировать в приближенных решениях, полученных тем и другим способами?

3.4. Докажите, что при любом начальном векторе $\left(x^{(0)}, y^{(0)}, z^{(0)}\right)^T$

последовательности векторов $\left(x_1^{(k)}, y_1^{(k)}, z_1^{(k)}\right)^T$ и $\left(x_2^{(k)}, y_2^{(k)}, z_2^{(k)}\right)^T$, определяемые при $k = 0, 1, 2, \ldots$ равенствами

$$\begin{cases} x_1^{(k+1)} = 0.1x_1^{(k)} + 0.2y_1^{(k)} - 3, \\ y_1^{(k+1)} = 0.2x_1^{(k)} - 0.1y_1^{(k)} + 0.1z_1^{(k)} + 2, \\ z_1^{(k+1)} = -0.3x_1^{(k)} + 0.2z_1^{(k)} - 1 \end{cases}$$

и

$$\begin{cases} x_2^{(k+1)} = \left(2y_2^{(k)} - 30\right)/9, \\ y_2^{(k+1)} = \left(2x_2^{(k)} + z_2^{(k)} + 20\right)/11, \\ z_2^{(k+1)} = -\left(3x_2^{(k)} + 10\right)/8, \end{cases}$$

сходятся, причем к одному и тому же предельному вектору $\left(x^*, y^*, z^*\right)^T$. Запишите линейную систему (в стандартном виде), решением которой служит этот предельный вектор. За сколько шагов итераций по данным формулам можно получить предельный вектор с точностью $\varepsilon = 10^{-6}$ (по норме-максимум), если начать счет с нулевого вектора?

3.5. Пусть методом Якоби решение системы

$$b_i x_{i-1} + c_i x_i + d_i x_{i+1} = r_i \quad (i = 1, 2, \ldots, n; \quad b_1 = d_n = 0)$$

с нужной точностью достигается за k шагов. Существуют ли такие k и n, при которых применение метода Якоби в этой ситуации эффективнее метода прогонки по числу арифметических операций?

3.6. Для линейной системы

$$\begin{cases} x_1 + 2x_2 + 3x_3 = 5, \\ 2x_1 - x_2 + 2x_4 = 8, \\ 3x_1 + x_2 - 2x_3 - x_4 = -1, \\ 4x_1 - x_3 + 2x_4 = 8 \end{cases}$$

запишите метод Зейделя и обоснуйте его сходимость. Каковы расчетные формулы метода ПВР в этом случае?

3.7. Убедитесь, что к системе

$$\begin{cases} x_1 + 2x_2 = 3, \\ 2x_1 + 2x_2 + x_3 = 5, \\ x_2 + 2x_3 = 3 \end{cases}$$

неприменим метод Якоби и что ее матрица не является положительно определенной. Выполните симметризацию Гаусса и убедитесь, что новая

система — нормальная. Запишите для нее процессы Зейделя и релаксации. Опробуйте последний при значениях параметра релаксации ω, больших и меньших единицы (например, полагая $\omega = 1.2$ и $\omega = 0.8$). Сравните результаты применения методов верхней, нижней и полной релаксации.

3.8. Дана система

$$\begin{cases} 7x_1 + 5x_2 + x_3 = 2.2, \\ 5x_1 + 8x_2 + 2x_3 = 2.4, \\ x_1 + 2x_2 + 4x_3 = 1.6. \end{cases}$$

Найдите четвертое приближение к ее решению по методу минимальных невязок, начиная итерационный процесс с нулевого вектора. За сколько итераций по методу Якоби достигается примерно такая же величина евклидовой нормы невязки? Сравните вычислительные затраты, требующиеся для реализации одного шага каждого из этих методов.

3.9. Методом сопряженных градиентов решите систему

$$\begin{cases} 3x_1 + 2x_2 + x_3 = 3, \\ x_1 + 2x_2 - x_3 = -2, \\ x_1 - x_2 + 2x_3 = 4. \end{cases}$$

3.10. Предположим, что некоторая $n \times n$-система вида $\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{c}$ с $\|\mathbf{B}\| \approx 0.5$ решается методом простых итераций с уровнем абсолютных погрешностей арифметических операций порядка 10^{-6}. Допустим, что при этом $\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\| \approx 1$. Каким числом следует ограничить количество итераций, чтобы вычислительная погрешность не стала существенно превышать погрешность метода?

3.11. Даны матрицы

$$\mathbf{A} = \begin{pmatrix} 1 & -2 & 3 \\ -1 & 1 & 2 \\ 2 & -1 & -1 \end{pmatrix} \quad \text{и} \quad \mathbf{U}_0 = \begin{pmatrix} -0.1 & 0.6 & 0.9 \\ -0.4 & 0.9 & 0.6 \\ 0.1 & 0.4 & 0.1 \end{pmatrix}.$$

А) Подсчитав невязку $\mathbf{\Psi}_0 = \mathbf{E} - \mathbf{A}\mathbf{U}_0$, убедитесь в существовании матрицы $\mathbf{A}^{-1}$ и оцените какую-либо ее норму.

Б) Сделайте по два приближения к $\mathbf{A}^{-1}$ методом Шульца второго и третьего порядков и оцените близость полученных приближений к $\mathbf{A}^{-1}$.

В) Сравните о ц е н к и погрешностей приближений с истинными ошибками, найдя $\mathbf{A}^{-1}$ каким-нибудь прямым методом.

ГЛАВА 4 МЕТОДЫ РЕШЕНИЯ АЛГЕБРАИЧЕСКИХ ПРОБЛЕМ СОБСТВЕННЫХ ЗНАЧЕНИЙ

Затрагивается наиболее сложная задача вычислительной линейной алгебры — нахождение собственных чисел и собственных векторов матриц. Рассматриваются современные подходы к решению спектральных задач для вещественных матриц умеренной размерности, базирующиеся на прямой и обратной итерациях (в том числе, со сдвигами), а также на приведении матриц к диагональной или треугольной формам ортогональными преобразованиями подобия. Показываются идеи методов, выводятся расчетные формулы, даются конкретные алгоритмы, позволяющие в оговоренных ситуациях решать частичные и полные проблемы собственных значений до конца.

4.1. СОБСТВЕННЫЕ ПАРЫ МАТРИЦЫ И ИХ ПРОСТЕЙШИЕ СВОЙСТВА

Пусть $\mathbf{A}$ — вещественная $n \times n$-матрица, $\mathbf{y} = \mathbf{y}(t)$ — n-мерная векторная функция скалярного аргумента t, и пусть ищутся нетривиальные решения системы дифференциальных уравнений

$$\frac{d\mathbf{y}}{dt} = \mathbf{A}\mathbf{y} \tag{4.1}$$

в виде $\mathbf{y} = e^{\lambda t}\mathbf{x}$, где $\mathbf{x} \in \mathbf{C}_n$, $\lambda \in \mathbf{C}$. Подставляя $\mathbf{y}$ и $\dfrac{d\mathbf{y}}{dt}$ в (4.1), получаем

$$\lambda e^{\lambda t}\mathbf{x} = \mathbf{A}e^{\lambda t}\mathbf{x},$$

т.е. система (4.1) действительно будет иметь решения заданного вида в том и только том случае, если найдутся такие пары чисел λ и ненулевых векторов $\mathbf{x}$, что

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}. \tag{4.2}$$

Имеется ряд других примеров из областей, лежащих за пределами линейной алгебры, в которых также приходят к необходимости решать подобные (4.2) алгебраические задачи, называемые ***задачами на собственные значения*** (см., например, [85]). При этом различают ***полную*** (***алгебраическую*** или, иначе,

матричную) проблему собственных значений, предполагающую нахождение всех ***собственных пар*** $\{\lambda, \mathbf{x}\}$ матрицы $\mathbf{A}$, и ***частичные проблемы собственных значений***, состоящие, как правило, в нахождении одного или нескольких ***собственных чисел*** λ и, возможно, соответствующих им ***собственных векторов****) $\mathbf{x}$. Чаще всего, в последнем случае речь идет о нахождении наибольшего и наименьшего по модулю собственных чисел; знание таких характеристик матрицы позволяет, например, делать заключения о сходимости тех или иных итерационных методов, оптимизировать параметры итерационных методов, учитывать влияние на результаты решения алгебраических задач погрешностей исходных данных и вычислительных погрешностей (потребность в таких числах неоднократно возникала в гл.3). Имеются и несколько иные постановки частичных проблем [12, 13, 44, 85].

Трактуя $\mathbf{A}$ в равенстве (4.2) как матрицу линейного преобразования в пространстве $\mathbf{R}_n$, задачу на собственные значения можно сформулировать так: для каких ненулевых векторов $\mathbf{x}$ и чисел λ линейное преобразование вектора с помощью матрицы $\mathbf{A}$ не изменяет направления этого вектора в $\mathbf{R}_n$, т.е. сводится к «растяжению» этого вектора в λ раз? Эта задача, очевидно, эквивалентна задаче исследования однородной СЛАУ**) с параметром: при каких λ система

$$(\mathbf{A} - \lambda \mathbf{E})\mathbf{x} = \mathbf{0} \tag{4.3}$$

имеет нетривиальные решения? Найти эти решения.

Теоретически эта задача легко решаема: нужно найти корни так называемого ***характеристического*** или иначе ***«векового»*** уравнения

$$\det(\mathbf{A} - \lambda \mathbf{E}) = 0 \tag{4.4}$$

и, подставляя их поочередно в (4.3), получать из соответствующих переопределенных систем собственные векторы. Практическая реализация этого в сущности простого подхода сопряжена с рядом трудностей, возрастающих с ростом размерности решаемой задачи. Трудности эти обусловлены развертыванием ***«векового» определителя*** $\det(\mathbf{A} - \lambda \mathbf{E})$ и вычислением корней получающегося при этом многочлена n-й степени, а также поиском линейно независимых решений вырожденных СЛАУ. В связи с этим такой непосредственный подход к решению алгебраической проблемы собственных значений обычно применяют лишь при очень малых размерах матриц $\mathbf{A}$ $(n = 2, 3)$; уже при $n \ge 4$ на первый план выходят специальные численные методы решения таких задач. Ниже будут рассмотрены некоторые из этих методов так, чтобы можно было понять идеи, лежащие в их основе, и в то же время получить возможность решать поставленные задачи до конца для некоторых классов матриц (более полное и глубокое изложение этой темы см. в монографиях [75, 141, 179, 180] и в учебных пособиях [3, 12, 13, 20, 42, 61, 78, 99].

Следует заметить, что если в недалеком прошлом численные методы решения задач на собственные значения опирались, как правило, на классический подход, т.е. на развертывание вековых определителей, в частности, в простейшем случае с помощью приведения матрицы $\mathbf{A}$ подходящим преобразованием к так называемой ***сопровождающей матрице*** [75, 179]

$$\mathbf{C} = \begin{pmatrix} c_1 & c_2 & c_3 & \dots & c_{n-1} & c_n \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix},$$

где в первой строке стоят коэффициенты уравнения (4.4), записанного в виде

$$(-1)^n \left(\lambda^n - c_1 \lambda^{n-1} - c_2 \lambda^{n-2} - \ldots - c_{n-1}\lambda - c_n\right) = 0,$$

то современные методы решения полной проблемы ориентированы на алгоритмическое построение из матрицы $\mathbf{A}$ такой матрицы, определенные элементы которой являлись бы приближенными значениями собственных чисел $\mathbf{A}$, причем параллельно формировались бы и ее собственные векторы.

Прежде чем приступить к изучению методов нахождения собственных чисел и векторов, вспомним некоторые простые их **свойства**, требующиеся в дальнейшем.

*) Собственное число и собственное значение в данном контексте — синонимы. В более общем случае, когда $\mathbf{A}$ в (4.2) — некоторый оператор, ***собственные элементы*** $\mathbf{x}$ могут иметь другую природу.

**) Здесь, как и ранее, аббревиатура СЛАУ означает «система линейных алгебраических уравнений», а обозначение $\mathbf{E}$ зарезервировано за единичной матрицей.

Свойство 4.1. *Если* $\{\lambda, \mathbf{x}\}$ *— собственная пара матрицы* $\mathbf{A}$, *а* $\alpha(\neq 0)$ *— некоторое число, то* $\{\lambda, \alpha\mathbf{x}\}$ *также является собственной парой для* $\mathbf{A}$.

Действительно, умножив верное для данных λ и $\mathbf{x}$ равенство (4.2) на число α, получаем верное равенство

$$\mathbf{A}(\alpha\mathbf{x}) = \lambda(\alpha\mathbf{x}).$$

Оно означает, что каждому собственному числу λ соответствует бесчисленное множество собственных векторов, различающихся лишь скалярным множителем. Такие векторы задают одно и то же направление в n-мерном пространстве; в соответствие этому направлению можно поставить нормированный вектор или орт (вообще говоря, одному собственному числу может соответствовать и несколько линейно независимых собственных векторов).

Свойство 4.2. *Пусть* $\{\mu, \mathbf{x}\}$ *— собственная пара матрицы* $\mathbf{A} - p\mathbf{E}$ *при некотором* $p \in \mathbf{R}$. *Тогда* $\{\lambda := \mu + p, \mathbf{x}\}$ *собственная пара матрицы* $\mathbf{A}$.

Чтобы убедиться в этом, заметим, что по условию

$$(\mathbf{A} - p\mathbf{E})\mathbf{x} = \mu\mathbf{x} \qquad (4.5)$$

при данных μ и $\mathbf{x}$ — верное равенство. Рассмотрим равенство $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ при $\lambda = \mu + p$:

$$\mathbf{A}\mathbf{x} = (\mu + p)\mathbf{x}.$$

Оно равносильно (4.5), и значит, справедливо, с другой стороны, говорит о том что $\{\lambda, \mathbf{x}\}$ — собственная пара $\mathbf{A}$.

Как видим, *прибавление к данной матрице* $\mathbf{A}$ *скалярной матрицы* $p\mathbf{E}$ *не изменяет ее собственных векторов и смещает спектр**) *исходной матрицы на число* p (влево при $p>0$).

Свойство 4.3. *Если* $\{\lambda, \mathbf{x}\}$ *— собственная пара обратимой матрицы* $\mathbf{A}$, *то* $\left\{\frac{1}{\lambda}, \mathbf{x}\right\}$ *— собственная пара матрицы* $\mathbf{A}^{-1}$.

Справедливость этого свойства очевидна: умножив верное

*) Напомним, что ***спектром матрицы*** называется множество всех ее собственных значений.

для данных λ и $\mathbf{x}$ равенство $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ слева на матрицу $\frac{1}{\lambda}\mathbf{A}^{-1}$, получаем

$$\frac{1}{\lambda}\mathbf{x} = \mathbf{A}^{-1}\mathbf{x},$$

что и означает утверждаемое.

Свойство 4.4. *Собственными числами диагональных и треугольных матриц являются их диагональные элементы.*

Этот факт легко усматривается из очевидного представления характеристических уравнений (4.4) для таких матриц в виде

$$\prod_{i=1}^{n}(\lambda - a_{ii}) = 0.$$

Последнее равенство свидетельствует о том, что *диагональные* и *треугольные вещественные матрицы имеют только вещественные собственные значения* (ровно n с учетом возможной их кратности). *Вещественность собственных чисел присуща и очень важному в приложениях классу симметричных матриц* [3, 43].

Определение 4.1. ***Отношением Рэлея****) *для* $n \times n$-*матрицы* $\mathbf{A}$ *называется функционал* $\rho(\mathbf{x}) = \dfrac{(\mathbf{A}\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})}$, *определенный на множестве ненулевых* n-*мерных векторов* $\mathbf{x}$.

Свойство 4.5. *Пусть* $\mathbf{x}^*$ *— собственный вектор матрицы* $\mathbf{A}$, *тогда* $\rho(\mathbf{x}^*)$ *— ее собственное число.*

Для доказательства этого утверждения обозначим через λ^* собственное число матрицы $\mathbf{A}$, соответствующее вектору $\mathbf{x}^*$. Подставляя $\mathbf{A}\mathbf{x}^* = \lambda^*\mathbf{x}^*$ в вытекающее из определения 4.1 равенство

$$(\mathbf{A}\mathbf{x}^*, \mathbf{x}^*) = \rho(\mathbf{x}^*)(\mathbf{x}^*, \mathbf{x}^*),$$

*) Лорд Рэлéй (до получения титула лорда — Стретт Джон Уильям, 1842–1919) — английский физик, один из основоположников теории колебаний.

имеем

$$\lambda^*(\mathbf{x}^*,\ \mathbf{x}^*) = \rho(\mathbf{x}^*)(\mathbf{x}^*,\ \mathbf{x}^*),$$

откуда после деления на $(\mathbf{x}^*,\ \mathbf{x}^*) \neq 0$ получаем утверждаемое: $\lambda^* = \rho(\mathbf{x}^*)$.

Отношение Рэлея обладает рядом других ценных свойств. Например, если матрица $\mathbf{A}$ — симметричная положительно определенная со спектром $\lambda_1 > \lambda_2 > \ldots > \lambda_n$, то $\max \rho(\mathbf{x}) = \lambda_1$, $\min \rho(\mathbf{x}) = \lambda_n$, $\rho(\mathbf{x}) \in [\lambda_n, \lambda_1]$ при любых n-мерных ненулевых векторах $\mathbf{x}$ и, кроме того, $\operatorname{grad} \rho(\mathbf{x}) = \mathbf{0}$ тогда и только тогда, когда $\mathbf{x}$ — собственный вектор матрицы $\mathbf{A}$ (см. [3, 141]). Эти свойства служат основой для некоторых способов локализации собственных значений и построения градиентных методов их вычисления.

В дальнейшем (§ 4.2, § 4.3) будет полезно следующее экстремальное свойство отношения Рэлея.

Свойство 4.6. *Минимум евклидовой нормы вектора $\xi(\lambda) := \mathbf{A}\mathbf{x} - \lambda\mathbf{x}$ для любого фиксированного ненулевого вектора $\mathbf{x}$ достигается при $\lambda = \rho(\mathbf{x})$.*

Смысл этого факта в следующем: *если некоторый вектор $\mathbf{x}$ считать приближением к собственному вектору матрицы $\mathbf{A}$* (а значит, ξ — его невязкой), *то отношение Рэлея $\rho(\mathbf{x})$ будет наилучшим приближением к соответствующему этому вектору собственному числу в смысле евклидовой метрики.*

Доказательство свойства для симметричных $\mathbf{A}$ и вещественных $\mathbf{x}$ весьма просто. Действительно, рассмотрим квадрат евклидовой нормы невязки

$$\|\xi(\lambda)\|_2^2 = (\mathbf{A}\mathbf{x} - \lambda\mathbf{x},\ \mathbf{A}\mathbf{x} - \lambda\mathbf{x}) =$$
$$= (\mathbf{A}\mathbf{x},\ \mathbf{A}\mathbf{x}) - 2\lambda(\mathbf{A}\mathbf{x},\ \mathbf{x}) + \lambda^2(\mathbf{x},\ \mathbf{x}) = q(\lambda)(\mathbf{x},\ \mathbf{x}),$$

где $q(\lambda) := \lambda^2 - 2\lambda\rho(\mathbf{x}) + \dfrac{(\mathbf{A}\mathbf{x},\ \mathbf{A}\mathbf{x})}{(\mathbf{x},\ \mathbf{x})}$. Очевидно, квадратный трехчлен $q(\lambda)$ всегда имеет минимум при $\lambda = \rho(\mathbf{x})$, а поскольку $(\mathbf{x},\ \mathbf{x}) > 0$, это значение λ доставляет минимум величине $\|\xi(\lambda)\|_2^2$, и следовательно, величине $\|\xi(\lambda)\|_2$.

4.2. СТЕПЕННОЙ МЕТОД

Рассмотрим простейший метод решения частичных проблем собственных значений, который вряд ли может быть отнесен к широко применяемым методам решения таких задач, но который много значит для понимания и построения других, более эффективных методов.

Пусть о вещественной $n \times n$-матрице $\mathbf{A}$ известно, что она является ***матрицей простой структуры***, т.е. имеет ровно n линейно независимых собственных векторов (базис):

$$\mathbf{x}_1 = \begin{pmatrix} x_{11} \\ x_{12} \\ \ldots \\ x_{1n} \end{pmatrix},\ \mathbf{x}_2 = \begin{pmatrix} x_{21} \\ x_{22} \\ \ldots \\ x_{2n} \end{pmatrix},\ \ldots,\ \mathbf{x}_n = \begin{pmatrix} x_{n1} \\ x_{n2} \\ \ldots \\ x_{nn} \end{pmatrix}. \qquad (4.6)$$

Пусть нумерация этих векторов отвечает упорядочению соответствующих им собственных чисел по убыванию модулей (где первое из неравенств — строгое):

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|. \qquad (4.7)$$

Ставим **задачу приближенного вычисления наибольшего по модулю собственного числа** λ_1 (вещественного, в силу предположения о строгом доминировании его модуля) **и соответствующего ему собственного вектора** $\mathbf{x}_1$ данной матрицы $\mathbf{A}$.

Возьмем произвольный ненулевой вектор $\mathbf{y}^{(0)}$ и запишем его разложение по базису из собственных векторов $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$:

$$\mathbf{y}^{(0)} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \ldots + c_n\mathbf{x}_n. \qquad (4.8)$$

При этом без ограничения общности можно считать, что $c_1 \neq 0$, так как в противном (маловероятном) случае можно взять другой начальный вектор $\mathbf{y}^{(0)}$.

Выполним первую итерацию вектора $\mathbf{y}^{(0)}$ умножением равенства (4.8) слева на матрицу $\mathbf{A}$:

$$\mathbf{y}^{(1)} = \mathbf{A}\mathbf{y}^{(0)} = c_1\mathbf{A}\mathbf{x}_1 + c_2\mathbf{A}\mathbf{x}_2 + \ldots + c_n\mathbf{A}\mathbf{x}_n.$$

Так как $\{\lambda_i, \mathbf{x}_i\}$ при всех $i \in \{1,\ 2,\ \ldots,\ n\}$ по предположению являются собственными парами матрицы $\mathbf{A}$, то, в силу (4.2), последнее можно переписать в виде

$$\mathbf{y}^{(1)} = c_1\lambda_1\mathbf{x}_1 + c_2\lambda_2\mathbf{x}_2 + \ldots + c_n\lambda_n\mathbf{x}_n.$$

Для второй итерации по тому же принципу получаем

$$\mathbf{y}^{(2)} = \mathbf{A}\mathbf{y}^{(1)} = \mathbf{A}^2\mathbf{y}^{(0)} = \\ = c_1\lambda_1\mathbf{A}\mathbf{x}_1 + c_2\lambda_2\mathbf{A}\mathbf{x}_2 + \ldots + c_n\lambda_n\mathbf{A}\mathbf{x}_n = \\ = c_1\lambda_1^2\mathbf{x}_1 + c_2\lambda_2^2\mathbf{x}_2 + \ldots + c_n\lambda_n^2\mathbf{x}_n.$$

Очевидно, k-я итерация вектора $\mathbf{y}^{(0)}$ с помощью матрицы $\mathbf{A}$ дает вектор

$$\mathbf{y}^{(k)} = \mathbf{A}\mathbf{y}^{(k-1)} = \mathbf{A}^k\mathbf{y}^{(0)} = c_1\lambda_1^k\mathbf{x}_1 + c_2\lambda_2^k\mathbf{x}_2 + \ldots + c_n\lambda_n^k\mathbf{x}_n \qquad (4.9)$$

или, с учетом представления $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ в исходном базисе (см. (4.6)),

$$\mathbf{y}^{(k)} = \begin{pmatrix} y_1^{(k)} \\ y_2^{(k)} \\ \ldots \\ y_n^{(k)} \end{pmatrix} = c_1\lambda_1^k \begin{pmatrix} x_{11} \\ x_{12} \\ \ldots \\ x_{1n} \end{pmatrix} + c_2\lambda_2^k \begin{pmatrix} x_{21} \\ x_{22} \\ \ldots \\ x_{2n} \end{pmatrix} + \ldots + c_n\lambda_n^k \begin{pmatrix} x_{n1} \\ x_{n2} \\ \ldots \\ x_{nn} \end{pmatrix}.$$

Беря отношения компонент ***итерированного вектора****) $\mathbf{y}^{(k)}$ к соответствующим компонентам предыдущего вектора $\mathbf{y}^{(k-1)}$, будем иметь:

$$\frac{y_i^{(k)}}{y_i^{(k-1)}} = \frac{c_1\lambda_1^k x_{1i} + c_2\lambda_2^k x_{2i} + \ldots + c_n\lambda_n^k x_{ni}}{c_1\lambda_1^{k-1} x_{1i} + c_2\lambda_2^{k-1} x_{2i} + \ldots + c_n\lambda_n^{k-1} x_{ni}} =$$

$$= \lambda_1 \cdot \frac{1 + \frac{c_2}{c_1}\cdot\frac{x_{2i}}{x_{1i}}\left(\frac{\lambda_2}{\lambda_1}\right)^k + \ldots + \frac{c_n}{c_1}\cdot\frac{x_{ni}}{x_{1i}}\left(\frac{\lambda_n}{\lambda_1}\right)^k}{1 + \frac{c_2}{c_1}\cdot\frac{x_{2i}}{x_{1i}}\left(\frac{\lambda_2}{\lambda_1}\right)^{k-1} + \ldots + \frac{c_n}{c_1}\cdot\frac{x_{ni}}{x_{1i}}\left(\frac{\lambda_n}{\lambda_1}\right)^{k-1}}. \qquad (4.10)$$

Предел дроби в последнем равенстве при сделанных допущениях равен 1 в процессе $k \to \infty$, и значит, $y_i^{(k)} \big/ y_i^{(k-1)} \xrightarrow[k\to\infty]{} \lambda_1$ для каждого $i \in \{1, 2, \ldots, n\}$, при котором $x_{1i} \neq 0$ (заметим, что числа $x_{11}, x_{12}, \ldots, x_{1n}$ не могут быть одновременно нулями, так как $\mathbf{x}_1$ — базисный вектор и поэтому не может быть нулевым).

---

*) Термин взят из [86].

Представляя вектор $\mathbf{y}^{(k)}$ на основе (4.9) в виде

$$\mathbf{y}^{(k)} = c_1\lambda_1^k\left[\mathbf{x}_1 + \frac{c_2}{c_1}\left(\frac{\lambda_2}{\lambda_1}\right)^k \mathbf{x}_2 + \ldots + \frac{c_n}{c_1}\left(\frac{\lambda_n}{\lambda_1}\right)^k \mathbf{x}_n\right], \qquad (4.11)$$

можно сделать вывод, что при тех же исходных допущениях, в силу $\left|\frac{\lambda_i}{\lambda_1}\right|^k \xrightarrow[\substack{k\to\infty \\ (i\neq 1)}]{} 0$, в фигурирующей в скобках выражения (4.11) линейной комбинации векторов $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ с ростом k начнет доминировать первое слагаемое. Это означает, что вектор $\mathbf{y}^{(k)}$ от итерации к итерации будет давать все более хорошие приближения к собственному вектору $\mathbf{x}_1$ по направлению, т.е. с точностью до скалярного множителя $c_1\lambda_1^k$ (см. свойство 4.1).

Таким образом, как показывают приведенные рассуждения, метод нахождения «старшей» собственной пары матрицы простой структуры, называемый ***степенным методом****), в своей основе весьма примитивен и состоит в следующем: берется произвольный вектор $\mathbf{y}^{(0)}\,(\neq \mathbf{0})$, простыми итерациями $\mathbf{y}^{(k)} = \mathbf{A}\mathbf{y}^{(k-1)}$ строится последовательность векторов $\mathbf{y}^{(k)}$ и параллельно рассматриваются последовательности отношений соответствующих компонент векторов k-й и $(k-1)$-й итераций (отношения с чрезвычайно малыми по модулю знаменателями следует игнорировать). Как только установятся несколько первых цифр во всех этих отношениях (что выясняется проверкой выполнения приближенных равенств $\frac{y_i^{(k)}}{y_i^{(k-1)}} \approx \frac{y_i^{(k-1)}}{y_i^{(k-2)}}$), так можно считать, что найдено наибольшее по модулю собственное число с точностью, определяемой последним установившимся в отношениях знаком, и соответствующий ему собственный вектор, за который принимается последний итерированный вектор $\mathbf{y}^{(k)}$.

Для практической реализации такая схема нахождения старшей собственной пары мало пригодна по многим причинам и требует определенной доводки. Рассмотрим некоторые из этих

---

*) Этимология данного термина совершенно ясна. Можно встретить и другие названия: ***счет на установление*** [78], ***итерационный метод фон Мизеса*** [87]. Иногда применяют латинскую аббревиатуру PM (от англ. *Power method*) [141].

причин и соответственно пути модификации вышеописанного простейшего алгоритма.

Анализируя выражение $\mathbf{y}^{(k)}$ в форме (4.11), видим, что при достаточно большом числе итераций k за счет множителя λ_1^k в процессе счета может произойти либо превышение допустимых для используемого компьютера чисел, если $|\lambda_1|>1$, либо пропадание значащих цифр итерированных векторов, если $|\lambda_1|<1$. Устранить это явление можно достаточно легко, введя в итерационный процесс нормировку итерированных векторов (т.е. приведение к единичной длине по той или иной метрике) на каждой итерации или через некоторое фиксированное число итерационных шагов.

Так, пошаговая нормировка векторов порождает следующий

РМ-алгоритм.

Шаг 1. Ввести $n \times n$-матрицу $\mathbf{A}$, задать n-мерный вектор $\mathbf{y}^{(0)}$, вычислить $\|\mathbf{y}^{(0)}\|$ и вектор $\mathbf{x}^{(0)} := \mathbf{y}^{(0)} / \|\mathbf{y}^{(0)}\|$; положить $k=1$.

Шаг 2. Вычислить вектор $\mathbf{y}^{(k)} = \mathbf{A}\mathbf{x}^{(k-1)}$.

Шаг 3. Вычислить $\|\mathbf{y}^{(k)}\|$ и $\mathbf{x}^{(k)} := \mathbf{y}^{(k)} / \|\mathbf{y}^{(k)}\|$.

Шаг 4. Вычислить отношения $\lambda_i^{(k)} = y_i^{(k)} / x_i^{(k-1)}$ (координат векторов $\mathbf{y}^{(k)}$ и $\mathbf{x}^{(k-1)}$) при $i \in \{1, 2, \ldots, n\}$ таких, что $|x_i^{(k-1)}| > \delta$, где $\delta > 0$ — некоторое задаваемое малое число (допуск).

Шаг 5. Подвергнуть числа $\lambda_i^{(k)}$ тесту на сходимость.

Если обнаруживается совпадение требуемого числа знаков в $\lambda_i^{(k)}$ и $\lambda_i^{(k-1)}$ ($\lambda^{(0)}$ можно задавать произвольно), то работу алгоритма прекратить и за старшее собственное число λ_1 принять усредненное (по i) значение $\lambda_i^{(k)}$, а за нормированный старший собственный вектор $\mathbf{x}_1$ — вектор $\mathbf{x}^{(k)}$.

В противном случае — вернуться к шагу 2.

Слабым местом данного алгоритма, очевидно, является последний шаг, т.е. решение проблемы своевременного останова работы алгоритма. Этот шаг описан из рациональных соображений и не может гарантировать во всех случаях (даже при сделанных допущениях) получения собственной пары $\{\lambda_1, \mathbf{x}_1\}$ с наперед

заданной точностью, поскольку при разработке метода не было получено никаких оценок погрешности.

Относительно характера сходимости степенного метода можно утверждать (см. формулы (4.10) и (4.11)), что в указанных условиях итерационный процесс является линейным*), т.е. сходится со скоростью геометрической прогрессии, знаменатель которой определяется в основном величиной отношения $\left|\frac{\lambda_2}{\lambda_1}\right|$. Это означает, что сходимость будет тем лучше и, как следствие, критерий останова в шаге 5 тем надежнее, чем сильнее доминирует в спектре матрицы $\mathbf{A}$ собственное число λ_1. Подмеченный факт вкупе со свойством 4.2 позволяет существенно ускорить нахождение наибольшего по модулю собственного числа матрицы $\mathbf{A}$ путем удачного смещения ее спектра, чему могут способствовать какие-либо априорные сведения об исходной задаче**).

То же свойство 4.2 собственных пар позволяет применять степенной метод непосредственно **для нахождения наименьшего по модулю собственного числа** λ_n знакоопределенной матрицы $\mathbf{A}$ в случае, когда наибольшее λ_1 уже найдено. Для этого достаточно найти наибольшее по модулю собственное число Λ матрицы $\mathbf{A} - \lambda_1 \mathbf{E}$; соответствующий ему собственный вектор этой матрицы и число $\lambda_n = \Lambda + \lambda_1$ будут образовывать искомую собственную пару.

Действительно, вычитая из верного для собственной пары

---

*) См. определение 5.1 в § 5.3.

) См. по этому поводу [179, 180]. В [179] приведен **пример, наглядно показывающий эффективность подходящего сдвига : если некая матрица $\mathbf{A}$ шестого порядка имеет собственные числа $\lambda_i = 21 - i$, то непосредственное применение степенного метода к вычислению λ_1 порождает итерационный процесс, сходящийся со скоростью порядка $\left(\frac{19}{20}\right)^k$, в то время как сдвиг на величину $p = 17$, оставляющий соответствующее λ_1 число $\mu_1 := \lambda_1 - p = 3$ старшим в спектре матрицы $\mathbf{A} - p\mathbf{E}$, позволяет найти его степенным методом, сходящимся уже со скоростью порядка $\left(\frac{2}{3}\right)^k$.

$\{\lambda_n, \mathbf{x}_n\}$ равенства $\mathbf{A}\mathbf{x}_n = \lambda_n \mathbf{x}_n$ тождество $\lambda_1 \mathbf{x}_n = \lambda_1 \mathbf{x}_n$, получаем верное равенство

$$(\mathbf{A} - \lambda_1 \mathbf{E})\mathbf{x}_n = (\lambda_n - \lambda_1)\mathbf{x}_n,$$

означающее, что $\Lambda := \lambda_n - \lambda_1$ и $\mathbf{x}_n$ служат собственной парой матрицы $\mathbf{A} - \lambda_1 \mathbf{E}$. Так как для знакоопределенной матрицы справедливо неравенство $|\lambda_n - \lambda_1| \geq |\lambda_i - \lambda_1|$ при любом $i \in \{1, 2, \ldots, n\}$, то Λ — наибольшее по модулю собственное число матрицы $\mathbf{A} - \lambda_1 \mathbf{E}$ и может быть найдено степенным методом.

Знание старшего собственного числа λ_1 матрицы $\mathbf{A}$ простой структуры, получаемого в процессе прямых итераций по формулам (4.9), (4.10), в предположении, что

$$|\lambda_1| > |\lambda_2| > |\lambda_3| \geq \ldots \geq |\lambda_n|,$$

позволяет без больших дополнительных затрат найти **приближенное значение второго по модулю собственного числа** λ_2. Это можно сделать по формуле

$$\lambda_2 \approx \frac{y_i^{(k+1)} - \lambda_1 y_i^{(k)}}{y_i^{(k)} - \lambda_1 y_i^{(k-1)}}, \tag{4.12}$$

вычисляя фигурирующие в правой части отношения для достаточно больших k и для всех $i \in \{1, 2, \ldots, n\}$, при которых абсолютная величина знаменателя не меньше некоторого порогового значения, и затем усредняя результат. Понятно, что при этом неизбежна потеря точности.

Для обоснования*) приближенного равенства (4.12) подставим в его правую часть выражения компонент $(k+1)$-го, k-го и $(k-1)$-го итерированных векторов в соответствии с представлением (4.9) в исходном базисе. После взаимного уничтожения по паре первых членов в числителе и в знаменателе будем иметь:

$$\frac{y_i^{(k+1)} - \lambda_1 y_i^{(k)}}{y_i^{(k)} - \lambda_1 y_i^{(k-1)}} = \frac{c_2 \lambda_2^{k+1} x_{2i} - c_2 \lambda_1 \lambda_2^k x_{2i} + \ldots + c_n \lambda_n^{k+1} x_{ni} - c_n \lambda_1 \lambda_n^k x_{ni}}{c_2 \lambda_2^k x_{2i} - c_2 \lambda_1 \lambda_2^{k-1} x_{2i} + \ldots + c_n \lambda_n^k x_{ni} - c_n \lambda_1 \lambda_n^{k-1} x_{ni}} =$$

$$= \frac{c_2 \lambda_2^{k+1} x_{2i} \left(1 - \dfrac{\lambda_1}{\lambda_2} + \sum\limits_{j=3}^{n} \dfrac{\lambda_j^{k+1} - \lambda_1 \lambda_j^k}{\lambda_2^{k+1}} \cdot \dfrac{c_j}{c_2} \cdot \dfrac{x_{ji}}{x_{2i}}\right)}{c_2 \lambda_2^k x_{2i} \left(1 - \dfrac{\lambda_1}{\lambda_2} + \sum\limits_{j=3}^{n} \dfrac{\lambda_j^k - \lambda_1 \lambda_j^{k-1}}{\lambda_2^k} \cdot \dfrac{c_j}{c_2} \cdot \dfrac{x_{ji}}{x_{2i}}\right)} \xrightarrow[k \to \infty]{} \lambda_2,$$

так как $\dfrac{\lambda_j^{k+1} - \lambda_1 \lambda_j^k}{\lambda_2^{k+1}} = \left(\dfrac{\lambda_j}{\lambda_2}\right)^{k+1} - \dfrac{\lambda_1}{\lambda_2} \cdot \left(\dfrac{\lambda_j}{\lambda_2}\right)^k \xrightarrow[k \to \infty]{} 0$ при всех $j = 3, \ldots, n$.

Вернемся к вопросу о недостатках степенного метода нахождения наибольшего по модулю собственного числа и путях их устранения. При этом **далее ограничимся рассмотрением класса симметричных положительно определенных матриц**. Известно, что такие матрицы имеют положительный вещественный спектр $\lambda_1, \lambda_2, \ldots, \lambda_n$, ортонормированный базис из собственных векторов $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ и, естественно, являются матрицами простой структуры.

Обсудим шаг 4 предложенного выше PM-алгоритма.

Вычисление на каждом итерационном шаге отношений всех пар соответствующих компонент векторов $\mathbf{x}$ и $\mathbf{y} = \mathbf{A}\mathbf{x}$, да еще с определенными проверками, при больших значениях n требует значительных вычислительных затрат, хотя и дает о старшем собственном числе λ_1 дополнительную информацию: как утверждается в [87], значение λ_1 заключено между наименьшим и наибольшим из этих отношений, т.е. имеются двусторонние оценки λ_1 на каждой итерации.

Чтобы упростить соответствующую шагу 4 PM-алгоритма процедуру, проведем следующие рассуждения.

Пусть $\mathbf{R}_n$ — евклидово пространство, $\mathbf{A}$ — симметричная положительно определенная матрица и последовательность итерированных векторов $\mathbf{y}^{(k)}$ строится, как и ранее, по формулам (4.9).

Рассмотрим скалярные произведения $\left(\mathbf{y}^{(k)}, \mathbf{y}^{(k)}\right)$ и $\left(\mathbf{y}^{(k)}, \mathbf{y}^{(k-1)}\right)$. Выполняя умножение правых частей (4.9) по пра-

*) Другое обоснование см. например, в [61]. Там же показано, что за соответствующий λ_2 собственный вектор $\mathbf{x}_2$ можно принять нормированный вектор $\mathbf{y}^{(k+1)} - \lambda_1 \mathbf{y}^{(k)}$.

вилам умножения многочленов и учитывая ортонормированность собственных векторов, т.е. условие $(\mathbf{x}_i, \mathbf{x}_j) = \delta_{ij}$ при $i, j \in \{1, 2, \ldots, n\}$, имеем:

$$\left(\mathbf{y}^{(k)},\ \mathbf{y}^{(k)}\right) = c_1^2 \lambda_1^{2k} + c_2^2 \lambda_2^{2k} + \ldots + c_n^2 \lambda_n^{2k},$$

$$\left(\mathbf{y}^{(k)},\ \mathbf{y}^{(k-1)}\right) = c_1^2 \lambda_1^{2k-1} + c_2^2 \lambda_2^{2k-1} + \ldots + c_n^2 \lambda_n^{2k-1}.$$

Отношение этих чисел

$$\frac{\left(\mathbf{y}^{(k)},\ \mathbf{y}^{(k)}\right)}{\left(\mathbf{y}^{(k)},\ \mathbf{y}^{(k-1)}\right)} = \lambda_1 \cdot \frac{1 + \left(\frac{c_2}{c_1}\right)^2 \left(\frac{\lambda_2}{\lambda_1}\right)^{2k} + \ldots + \left(\frac{c_n}{c_1}\right)^2 \left(\frac{\lambda_n}{\lambda_1}\right)^{2k}}{1 + \left(\frac{c_2}{c_1}\right)^2 \left(\frac{\lambda_2}{\lambda_1}\right)^{2k-1} + \ldots + \left(\frac{c_n}{c_1}\right)^2 \left(\frac{\lambda_n}{\lambda_1}\right)^{2k-1}} \quad (4.13)$$

в оговоренных выше условиях при $k \to \infty$ имеет пределом наибольшее собственное число λ_1, причем скорость сходимости к пределу будет больше, чем в степенном методе, опирающемся на отношения (4.10) $\left(O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\right) \text{ против } O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{k}\right) \right)$.

Базирующаяся на таком подходе модификация степенного метода называется ***методом скалярных произведений***. Реализовать ее можно, например, в виде следующего ***SP-алгоритма****).

Шаг 1. Ввести: данную симметричную $n \times n$-матрицу $\mathbf{A}$, произвольный n-мерный начальный вектор $\mathbf{y}^{(0)} (\neq \mathbf{0})$, малое число $\varepsilon > 0$ (определяющее допустимую абсолютную погрешность искомого собственного числа λ_1), число $\lambda^{(0)}$ для начального сравнения (например, 0). Положить $k = 1$ (включить счетчик итераций).

Шаг 2. Вычислить скаляры $s^{(0)} = \left(\mathbf{y}^{(0)},\ \mathbf{y}^{(0)}\right)$, $\left\|\mathbf{y}^{(0)}\right\|_2 = \sqrt{s^{(0)}}$ и вектор $\mathbf{x}^{(0)} = \mathbf{y}^{(0)} / \left\|\mathbf{y}^{(0)}\right\|_2$.

Шаг 3. Вычислить $\mathbf{y}^{(k)} = \mathbf{A}\mathbf{x}^{(k-1)}$ (итерация нормированного вектора).

*) SP от англ. *Scalar product.*

Шаг 4. Вычислить: $s^{(k)} = \left(\mathbf{y}^{(k)},\ \mathbf{y}^{(k)}\right)$ и $t^{(k)} = \left(\mathbf{y}^{(k)},\ \mathbf{x}^{(k-1)}\right)$ (скалярные произведения), $\left\|\mathbf{y}^{(k)}\right\|_2 = \sqrt{s^{(k)}}$, $\mathbf{x}^{(k)} = \mathbf{y}^{(k)} / \left\|\mathbf{y}^{(k)}\right\|_2$ (приближение к нормированному собственному вектору), $\lambda^{(k)} = s^{(k)} / t^{(k)}$ (приближение к собственному числу λ_1).

Шаг 5. Если $\left|\lambda^{(k)} - \lambda^{(k-1)}\right| > \varepsilon$, положить $k := k+1$ и вернуться к шагу 3, иначе завершить работу алгоритма, считая $\lambda_1 :\approx \lambda^{(k)}$, $\mathbf{x}_1 :\approx \mathbf{x}^{(k)}$.

Замечание 4.1. Данный алгоритм позволяет более быстро (т.е. за меньшее число итераций), чем PM-алгоритм, найти с нужной точностью наибольшее собственное число с и м м е т р и ч н о й матрицы, но при этом точность приближенного равенства $\mathbf{x}_1 :\approx \mathbf{x}^{(k)}$ для соответствующего собственного вектора может оказаться недостаточной (объясните, почему?).

Замечание 4.2. Очевидно, в методе скалярных произведений вместо отношения (4.13), стремящегося к λ_1 при $k \to \infty$, можно с тем же успехом взять отношение $\left(\mathbf{y}^{(k+1)},\ \mathbf{y}^{(k)}\right)$ к $\left(\mathbf{y}^{(k)},\ \mathbf{y}^{(k)}\right)$, имеющее тот же предел. Последнее же есть не что иное, как отношение Рэлея:

$$\frac{\left(\mathbf{y}^{(k+1)},\ \mathbf{y}^{(k)}\right)}{\left(\mathbf{y}^{(k)},\ \mathbf{y}^{(k)}\right)} = \frac{\left(\mathbf{A}\mathbf{y}^{(k)},\ \mathbf{y}^{(k)}\right)}{\left(\mathbf{y}^{(k)},\ \mathbf{y}^{(k)}\right)} = \rho\left(\mathbf{y}^{(k)}\right);$$

отсюда другое название метода скалярных произведений — ***метод частных Рэлея***. В соответствии со свойствами 4.5, 4.6 предыдущего параграфа можно сказать, что этим методом на каждом итерационном шаге ищется наилучшее для вычисленного вектора $\mathbf{y}^{(k)}$ приближение к собственному числу λ_1 в смысле евклидовой нормы невязки.

Пример 4.1. На матрице $\mathbf{A} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ покажем п р о ц е с с построения приближений к старшему собственному числу λ_1 (и соответствующему ему собственному вектору $\mathbf{x}_1$) методом скалярных произведений.

Приняв за начальный вектор $\mathbf{y}^{(0)}$ первый орт $\mathbf{e}_1 = (1;\ 0)^T$, далее идем по SP-алгоритму. Имеем:

$$s^{(0)} = \left(\mathbf{y}^{(0)},\ \mathbf{y}^{(0)}\right) = 1, \quad \left\|\mathbf{y}^{(0)}\right\|_2 = \sqrt{s^{(0)}} = 1, \quad \mathbf{x}^{(0)} = \frac{\mathbf{y}^{(0)}}{\left\|\mathbf{y}^{(0)}\right\|_2} = \begin{pmatrix} 1 \\ 0 \end{pmatrix};$$

1-я итерация:

$$\mathbf{y}^{(1)} = \mathbf{A}\,\mathbf{x}^{(0)} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}, \quad s^{(1)} = \left(\mathbf{y}^{(1)},\ \mathbf{y}^{(1)}\right) = 5,$$

$$t^{(1)} = \left(\mathbf{y}^{(1)}, \mathbf{x}^{(0)}\right) = 2, \quad \left\|\mathbf{y}^{(1)}\right\|_2 = \sqrt{s^{(1)}} = \sqrt{5},$$

$$\mathbf{x}^{(1)} = \frac{\mathbf{y}^{(1)}}{\left\|\mathbf{y}^{(1)}\right\|_2} = \frac{1}{\sqrt{5}}\begin{pmatrix} 2 \\ -1 \end{pmatrix} \approx \begin{pmatrix} 0.894 \\ -0.447 \end{pmatrix}, \quad \lambda^{(1)} = \frac{s^{(1)}}{t^{(1)}} = \frac{5}{2} = 2.5;$$

2-я итерация:

$$\mathbf{y}^{(2)} = \mathbf{A}\,\mathbf{x}^{(1)} = \frac{1}{\sqrt{5}}\begin{pmatrix} 5 \\ -4 \end{pmatrix}, \quad s^{(2)} = \left(\mathbf{y}^{(2)},\ \mathbf{y}^{(2)}\right) = \frac{41}{5},$$

$$t^{(2)} = \left(\mathbf{y}^{(2)}, \mathbf{x}^{(1)}\right) = \frac{14}{5}, \quad \left\|\mathbf{y}^{(2)}\right\|_2 = \sqrt{s^{(2)}} = \sqrt{\frac{41}{5}},$$

$$\mathbf{x}^{(2)} = \frac{\mathbf{y}^{(2)}}{\left\|\mathbf{y}^{(2)}\right\|_2} = \frac{1}{\sqrt{41}}\begin{pmatrix} 5 \\ -4 \end{pmatrix} \approx \begin{pmatrix} 0.781 \\ -0.625 \end{pmatrix}, \quad \lambda^{(2)} = \frac{s^{(2)}}{t^{(2)}} = \frac{41}{14} \approx 2.929;$$

3-я итерация:

$$\mathbf{y}^{(3)} = \mathbf{A}\,\mathbf{x}^{(2)} = \frac{1}{\sqrt{41}}\begin{pmatrix} 14 \\ -13 \end{pmatrix}, \quad s^{(3)} = \left(\mathbf{y}^{(3)},\ \mathbf{y}^{(3)}\right) = \frac{365}{41},$$

$$t^{(3)} = \left(\mathbf{y}^{(3)}, \mathbf{x}^{(2)}\right) = \frac{122}{41}, \quad \left\|\mathbf{y}^{(3)}\right\|_2 = \sqrt{s^{(3)}} = \sqrt{\frac{365}{41}},$$

$$\mathbf{x}^{(3)} = \frac{\mathbf{y}^{(3)}}{\left\|\mathbf{y}^{(3)}\right\|_2} = \frac{1}{\sqrt{365}}\begin{pmatrix} 14 \\ -13 \end{pmatrix} \approx \begin{pmatrix} 0.733 \\ -0.680 \end{pmatrix}, \quad \lambda^{(3)} = \frac{s^{(3)}}{t^{(3)}} = \frac{365}{122} \approx 2.992.$$

По значениям величин $\left|\lambda^{(2)} - \lambda^{(1)}\right| \approx 0.429$, $\left|\lambda^{(3)} - \lambda^{(2)}\right| \approx 0.063$ можно судить о сближении последовательных приближений $\lambda^{(1)}$, $\lambda^{(2)}$, $\lambda^{(3)}$ собственного числа λ_1 с каждой итерацией. Последнюю из этих величин можно считать нестрогой оценкой абсолютной погрешности равенства $\lambda_1 \approx \lambda^{(3)}$ (на самом деле знание $\lambda_1 = 3$ показывает его более высокую точность: $\left|\lambda_1 - \lambda^{(3)}\right| \approx 0.008 < 0.01$). Для собственного вектора $\mathbf{x}_1$ приближенное равенство $\mathbf{x}_1 \approx \mathbf{x}^{(3)}$ можно оценить величиной

$$\left\|\mathbf{x}^{(3)} - \mathbf{x}^{(2)}\right\| \approx \left\|\begin{pmatrix} -0.048 \\ -0.055 \end{pmatrix}\right\|_2 \approx 0.073.$$

Наличие ортонормированного базиса из собственных векторов $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ матрицы $\mathbf{A}$ позволяет применять степенной метод (метод скалярных произведений) для последовательного вычисления собственных пар $\{\lambda_i, \mathbf{x}_i\}$ при $i \geq 2$ более совершенными, чем определяемый формулой (4.12), способами. Рассмотрим один из них.

Пусть первая (старшая) собственная пара $\{\lambda_1, \mathbf{x}_1\}$ уже найдена, причем $\|\mathbf{x}_1\| = \sqrt{(\mathbf{x}_1, \mathbf{x}_1)} = 1$. Возьмем произвольный ненулевой вектор $\mathbf{z}^{(0)}$ и образуем вектор

$$\mathbf{y}^{(0)} = \mathbf{z}^{(0)} - \left(\mathbf{z}^{(0)}, \mathbf{x}_1\right)\mathbf{x}_1. \qquad (4.14)$$

Так как

$$\left(\mathbf{y}^{(0)}, \mathbf{x}_1\right) = \left(\mathbf{z}^{(0)}, \mathbf{x}_1\right) - \left(\mathbf{z}^{(0)}, \mathbf{x}_1\right)(\mathbf{x}_1, \mathbf{x}_1) = 0,$$

то вектор $\mathbf{y}^{(0)}$ ортогонален $\mathbf{x}_1$, т.е. его проекция на первый базисный вектор системы $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ равна нулю. Значит, разложение (4.8) вектора $\mathbf{y}^{(0)}$ по этому базису имеет вид

$$\mathbf{y}^{(0)} = c_2\mathbf{x}_2 + c_3\mathbf{x}_3 + \ldots + c_n\mathbf{x}_n,$$

и, соответственно, степенные итерации этого вектора типа (4.9) порождают векторы

$$\mathbf{y}^{(k)} = c_2\lambda_2^k\mathbf{x}_2 + c_3\lambda_3^k\mathbf{x}_3 + \ldots + c_n\lambda_n^k\mathbf{x}_n. \qquad (4.15)$$

Легко видеть (сравните с (4.13)), что если $|\lambda_2| > |\lambda_i|$ при всех $i \in \{3, \ldots, n\}$, то $\dfrac{\left(\mathbf{y}^{(k)},\ \mathbf{y}^{(k)}\right)}{\left(\mathbf{y}^{(k)},\ \mathbf{y}^{(k-1)}\right)} \xrightarrow[k\to\infty]{} \lambda_2$ со скоростью $O\left(\left|\dfrac{\lambda_3}{\lambda_2}\right|^{2k}\right)$ и $\mathbf{x}^{(k)} = \dfrac{\mathbf{y}^{(k)}}{\left\|\mathbf{y}^{(k)}\right\|} \xrightarrow[k\to\infty]{} \mathbf{x}_2$ со скоростью $O\left(\left|\dfrac{\lambda_3}{\lambda_2}\right|^{k}\right)$.

Следующая собственная пара $\{\lambda_3, \mathbf{x}_3\}$ может быть найдена приближенно тем же методом, если за начальный вектор последовательности $\left(\mathbf{y}^{(k)}\right)$ принять вектор

$$\mathbf{y}^{(0)} = \mathbf{z}^{(0)} - \left(\mathbf{z}^{(0)}, \mathbf{x}_1\right)\mathbf{x}_1 - \left(\mathbf{z}^{(0)}, \mathbf{x}_2\right)\mathbf{x}_2,$$

ортогональный одновременно $\mathbf{x}_1$ и $\mathbf{x}_2$ при любом $\mathbf{z}^{(0)}$, и т.д.

Известны и другие способы последовательного нахождения собственных пар, опирающиеся на непосредственное применение степенного метода. При этом имеются возможности понижения размерности при нахождении каждой последующей собственной пары.

Замечание 4.3. В реальных расчетах, в силу неизбежных ошибок округлений, в представлении (4.15) итерированного вектора $\mathbf{y}^{(k)}$ при вычислении второй собственной пары появится малое, но растущее с увеличением номера k слагаемое, соответствующее проекции $\mathbf{y}^{(k)}$ на первый собственный вектор $\mathbf{x}_1$. Поэтому реальный алгоритм должен предусматривать возврат к началу процесса итерирования, т.е. проведение операции ортогонализации по формуле (4.14) с $\mathbf{z}^{(0)} := \mathbf{y}^{(m)}$ через некоторое число итераций $k = m$.

Замечание 4.4. Не всегда бывает известным, выполняются ли оговоренные выше условия, при которых изучался степенной метод. В таких ситуациях при его применении нужно принимать особые меры предосторожности. Целесообразно, например, контролировать, сближаются ли члены последовательности $(\lambda^{(k)})$ посредством проверки неравенств

$$\left|\lambda^{(k+1)} - \lambda^{(k)}\right| < \left|\lambda^{(k)} - \lambda^{(k-1)}\right|$$

(***прием Гарвика*** [3, 78]), а также осуществлять итерационный процесс с разных начальных векторов (в случае кратности находимого собственного числа это просто необходимо для вычисления степенным методом всех соответствующих ему собственных векторов).

Замечание 4.5. Как отмечалось выше, степенной метод сходится линейно, точнее, имеет лишь асимптотическую скорость сходимости геометрической прогрессии. При слабом доминировании модуля вычисляемого собственного числа эта сходимость может оказаться чрезвычайно медленной. Ускорения итерационного процесса можно достигнуть за счет быстрого накопления степеней матриц по схеме

$$\mathbf{A}\cdot\mathbf{A}=\mathbf{A}^2, \quad \mathbf{A}^2\cdot\mathbf{A}^2=\mathbf{A}^4$$

и т.д., что позволяет производить не последовательное, пошаговое, а скачкообразное построение последовательности $(\mathbf{y}^{(k)})$ с помощью равенств вида

$$\mathbf{y}^{(k)} = \mathbf{A}^{k-m}\mathbf{y}^{(m)}$$

при фиксированных $m \in \{0, 1, \ldots, k-1\}$ (таких, при которых $k-m$ является некоторой целой степенью двойки) и нормированием сразу после очередного скачка. Здесь, правда, нужно особенно внимательно относиться к риску выхода за границы диапазона компьютерных чисел в процессе счета.

Если не требуется находить собственный вектор $\mathbf{x}_1$, то более быстро вычислять максимальное по модулю собственное число λ_1 можно на основе соотношения [61]

$$\lambda_1^k + \lambda_2^k + \ldots + \lambda_n^k = \operatorname{Sp}\mathbf{A}^k \quad \forall k \in \mathbf{N}.$$

Вычислив $\mathbf{A}^k$ по закону удвоения степеней, а затем $\mathbf{A}^{k+1} = \mathbf{A}^k \cdot \mathbf{A}$, находим отношение следов (сумм диагональных элементов) этих матриц:

$$\frac{\operatorname{Sp}\mathbf{A}^{k+1}}{\operatorname{Sp}\mathbf{A}^k} = \frac{\lambda_1^{k+1}\left(1+\left(\frac{\lambda_2}{\lambda_1}\right)^{k+1}+\ldots+\left(\frac{\lambda_n}{\lambda_1}\right)^{k+1}\right)}{\lambda_1^k\left(1+\left(\frac{\lambda_2}{\lambda_1}\right)^{k}+\ldots+\left(\frac{\lambda_n}{\lambda_1}\right)^{k}\right)} \xrightarrow[k\to\infty]{} \lambda_1.$$

Замечание 4.6. Более популярный способ улучшения сходимости степенного метода — это применение Δ^2-процесса Эйткена *). Считается, что если $\lambda^{(k-1)}$, $\lambda^{(k)}$, $\lambda^{(k+1)}$ являются тремя последовательными приближениями к собственному числу, полученными степенным методом, то число

$$\tilde{\lambda} := \lambda^{(k-1)} - \frac{\left(\lambda^{(k)} - \lambda^{(k-1)}\right)^2}{\lambda^{(k+1)} - 2\lambda^{(k)} + \lambda^{(k-1)}}$$

ближе к пределу этой последовательности, чем каждое из них. Этот факт может быть использован в реальных алгоритмах либо через несколько итерационных шагов (например, через два на третий), либо на завершающем этапе вычислений. Для искомого собственного вектора такое ускорение может производиться п о к о о р д и н а т н о.

4.3. ОБРАТНЫЕ ИТЕРАЦИИ

В предыдущем параграфе было показано, что при определенных условиях наименьшее по модулю собственное число λ_n может быть найдено степенным методом, когда уже известно наибольшее λ_1. Если же проблема состоит в нахождении лишь младшей собственной пары матрицы $\mathbf{A}$, то можно обойтись и без вычисления λ_1, применяя степенной метод к матрице $\mathbf{A}^{-1}$.

В самом деле, если данная матрица $\mathbf{A}$ имеет собственные пары

$$\{\lambda_1, \mathbf{x}_1\}, \{\lambda_2, \mathbf{x}_2\}, \ldots, \{\lambda_{n-1}, \mathbf{x}_{n-1}\}, \{\lambda_n, \mathbf{x}_n\},$$

то по свойству 4.3 собственными парами матрицы $\mathbf{A}^{-1}$ будут

$$\left\{\frac{1}{\lambda_1}, \mathbf{x}_1\right\}, \left\{\frac{1}{\lambda_2}, \mathbf{x}_2\right\}, \ldots, \left\{\frac{1}{\lambda_{n-1}}, \mathbf{x}_{n-1}\right\}, \left\{\frac{1}{\lambda_n}, \mathbf{x}_n\right\}.$$

*) Более подробно об этом процессе см. в § 6.2.1.

При этом упорядочиванию спектра **A**

$$|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_{n-1}| > |\lambda_n|$$

соответствует цепочка неравенств

$$\left|\frac{1}{\lambda_n}\right| > \left|\frac{1}{\lambda_{n-1}}\right| \geq \ldots \geq \left|\frac{1}{\lambda_2}\right| \geq \left|\frac{1}{\lambda_1}\right|$$

для собственных чисел $\gamma_1 := \frac{1}{\lambda_n}$, $\gamma_2 := \frac{1}{\lambda_{n-1}}$, …, $\gamma_{n-1} := \frac{1}{\lambda_2}$, $\gamma_n := \frac{1}{\lambda_1}$ матрицы $\mathbf{A}^{-1}$. Это значит, что наименьшим по модулю собственным числом данной матрицы **A** является величина, обратная наибольшему по модулю собственному числу матрицы $\mathbf{A}^{-1}$. Последнее же может быть получено прямыми итерациями произвольного начального вектора $\mathbf{y}^{(0)}$ посредством матрицы $\mathbf{A}^{-1}$ по аналогичной (4.9) формуле

$$\mathbf{y}^{(k)} = \mathbf{A}^{-1}\mathbf{y}^{(k-1)}, \quad k = 1, 2, \ldots \tag{4.16}$$

При достаточно больших $k \in \mathbf{N}$ последовательность отношений одноименных координат векторов $\mathbf{y}^{(k)}$ и $\mathbf{y}^{(k-1)}$ должна давать приближенное значение $\frac{1}{\lambda_n}$, а вектор $\mathbf{y}^{(k)}$ (желательно его нормирование) можно принять за собственный вектор $\mathbf{x}_n$.

Вместо прямых итераций (4.16), требующих предварительного обращения исходной матрицы **A**, обычно предпочитают строить ту же последовательность векторов $\left(\mathbf{y}^{(k)}\right)$, решая при $k = 1, 2, 3, \ldots$ линейные системы

$$\mathbf{A}\mathbf{y}^{(k)} = \mathbf{y}^{(k-1)}. \tag{4.17}$$

Так как все эти системы имеют одну и ту же матрицу коэффициентов, то самая трудоемкая часть метода Гаусса для их решения — LU-факторизация матрицы **A** — может быть выполнена лишь один раз.

Построение последовательности векторов, приближающих собственный вектор $\mathbf{x}_n$ по неявной формуле (4.17), называют ***обратными итерациями***, а процесс решения частичных проблем собственных значений на этой основе — ***методом***

обратных итераций*).

Применение обратных итераций к нахождению младшей собственной пары матрицы **A** не требует написания специального алгоритма, достаточно лишь заменить один шаг в алгоритмах предыдущего параграфа. А именно, наполнение шага 2 в **PM**-алгоритме для матриц простой структуры и шага 3 в **SP**-алгоритме для симметричных матриц должно быть следующим:

решить уравнение $\mathbf{A}\mathbf{y}^{(k)} = \mathbf{y}^{(k-1)}$.

Полученный алгоритм называют ***INVIT-алгоритмом*****) [141].

Метод обратных итераций, а точнее, ***обратные итерации со сдвигами*** часто применяют в тех случаях, когда нужно с большой точностью найти собственный вектор, отвечающий какому-либо собственному числу из спектра заданной матрицы при условии, что известно приближенное значение этого числа. При этом, очевидно, прямое решение однородной системы (4.3) заведомо неприменимо, так как подстановка в нее значения λ, хоть сколько-нибудь отличного от собственного, сделает систему однозначно разрешимой, т.е. допускающей только тривиальное решение. Рассмотрим суть обратных итераций со сдвигами.

Пусть для собственного числа λ_j матрицы простой структуры **A** известно его приближение σ такое, что

$$|\lambda_j - \sigma| < |\lambda_i - \sigma| \quad \forall\, i \neq j, \tag{4.18}$$

т.е. число σ ближе к собственному числу λ_j, чем к какому-либо другому собственному числу матрицы **A**.

Начиная с вектора $\mathbf{x}^{(0)}$ такого, что $\left\|\mathbf{x}^{(0)}\right\| = 1$, образуем последовательность нормированных векторов $\left(\mathbf{x}^{(k)}\right)$ по формулам

$$(\mathbf{A} - \sigma\mathbf{E})\mathbf{y}^{(k)} = \mathbf{x}^{(k-1)}; \tag{4.19}$$

$$\mathbf{x}^{(k)} = \mathbf{y}^{(k)} / \left\|\mathbf{y}^{(k)}\right\|, \quad k = 1, 2, \ldots. \tag{4.20}$$

Изучим поведение этой последовательности, для чего запишем разложение векторов $\mathbf{y}^{(k)}$ и $\mathbf{x}^{(k-1)}$ по базису из собственных

---

*) Другое название ***обратный степенной метод*** [29].

**) От англ. *Inverse iteration*.

векторов $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ с некоторыми коэффициентами $c_i^{(k)}$ и $b_i^{(k-1)}$ соответственно:

$$\mathbf{y}^{(k)} = c_1^{(k)}\mathbf{x}_1 + c_2^{(k)}\mathbf{x}_2 + \ldots + c_n^{(k)}\mathbf{x}_n, \qquad (4.21)$$
$$\mathbf{x}^{(k-1)} = b_1^{(k-1)}\mathbf{x}_1 + b_2^{(k-1)}\mathbf{x}_2 + \ldots + b_n^{(k-1)}\mathbf{x}_n.$$

Подставляя это в (4.19) и учитывая, что по определению

$$\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i \quad \forall i \in \{1, 2, \ldots, n\},$$

имеем

$$(\lambda_1 - \sigma)c_1^{(k)}\mathbf{x}_1 + (\lambda_2 - \sigma)c_2^{(k)}\mathbf{x}_2 + \ldots + (\lambda_n - \sigma)c_n^{(k)}\mathbf{x}_n =$$
$$= b_1^{(k-1)}\mathbf{x}_1 + b_2^{(k-1)}\mathbf{x}_2 + \ldots + b_n^{(k-1)}\mathbf{x}_n,$$

откуда, в силу единственности разложения вектора по базису, следует

$$(\lambda_i - \sigma)c_i^{(k)} = b_i^{(k-1)} \quad \forall i \in \{1, 2, \ldots, n\}.$$

Анализируя получающиеся отсюда выражения

$$c_i^{(k)} = \frac{b_i^{(k-1)}}{\lambda_i - \sigma} \qquad (4.22)$$

коэффициентов разложения вектора $\mathbf{y}^{(k)}$ по базису из собственных векторов, видим, что вследствие малости модуля знаменателя $\lambda_j - \sigma$ по сравнению с другими знаменателями $\lambda_i - \sigma$ (см. (4.18)), можно рассчитывать на преимущественное возрастание коэффициентов $c_j^{(k)}$ именно при собственном векторе $\mathbf{x}_j$ с ростом k. Значит, чем сильнее неравенство в (4.18), тем сильнее (быстрее) будет доминировать составляющая собственного вектора $\mathbf{x}_j$ в представлении (4.21) вектора $\mathbf{y}^{(k)}$, а значит, и вектора $\mathbf{x}^{(k)}$, получаемого из $\mathbf{y}^{(k)}$ нормированием (4.20). Последнее же говорит о том, что каков бы ни был начальный вектор $\mathbf{x}^{(0)}$ $(\neq \mathbf{0})$*), быстрое доминирование $c_j^{(k)}$ среди остальных коэффициентов $c_i^{(k)}$ происходит еще и за счет числителей дробей (4.22).

---

*) Лишь бы при его выборе не попасть на ортогональный $\mathbf{x}_j$ вектор; зачастую берут вектор $\mathbf{x}^{(0)}$ с равными координатами.

Следует заметить, что обратные итерации со сдвигами (4.19), (4.20) позволяют не только найти собственный вектор $\mathbf{x}_j$, но и служат основой для уточнения приближенного равенства $\lambda_j \approx \sigma$.

Действительно, формулы (4.19), (4.20) определяют не что иное, как метод обратных итераций для нахождения наименьшего по модулю собственного числа матрицы $\mathbf{A} - \sigma\mathbf{E}$, и, если σ существенно ближе к λ_j, чем к любому другому собственному числу λ_i матрицы $\mathbf{A}$, то уточняющие λ_j значения, согласно свойству 4.2, можно получать при $k = 1, 2, \ldots$ по формуле

$$\lambda_j^{(k)} = \sigma + \left\langle \frac{x_i^{(k-1)}}{y_i^{(k)}} \right\rangle, \qquad (4.23)$$

где $x_i^{(k-1)}$ и $y_i^{(k)}$ — координаты векторов $\mathbf{x}^{(k-1)}$ и $\mathbf{y}^{(k)}$ соответственно, а $\langle \cdot \rangle$ — знак усреднения по всем $i \in \{1, 2, \ldots, n\}$, при которых $y_i^{(k)} \neq 0$ [78].

Как показывает практика вычислений, сходимость процесса обратных итераций со сдвигами характеризуется высокой скоростью (по сравнению с обычным степенным методом). Но еще более быстрая сходимость может быть получена введением переменных сдвигов, определяемых какой-нибудь последовательностью чисел $\sigma_0, \sigma_1, \sigma_2, \ldots$, сходящейся к находимому собственному числу. Не вызывает сомнений целесообразность использования в роли таких чисел приближений $\lambda_j^{(k)}$ к собственному числу λ_j, получаемых по формуле (4.23).

Таким образом, ***обратные итерации с переменными сдвигами*** можно определить совокупностью равенств

$$\left(\mathbf{A} - \lambda_j^{(k-1)}\mathbf{E}\right)\mathbf{y}^{(k)} = \mathbf{x}^{(k-1)},$$
$$\mathbf{x}^{(k)} = \mathbf{y}^{(k)} \big/ \left\|\mathbf{y}^{(k)}\right\|, \qquad (4.24)$$
$$\lambda_j^{(k)} = \lambda_j^{(k-1)} + \left\langle \frac{x_i^{(k-1)}}{y_i^{(k)}} \right\rangle,$$

где $k = 1, 2, \ldots$, а число $\lambda_j^{(0)}$ $(\approx \lambda_j)$ и вектор $\mathbf{x}^{(0)}$ (такой, что $\left\|\mathbf{x}^{(0)}\right\| = 1$) задаются.

Скорость сходимости процесса (4.24) — квадратичная [43, 78], в то время как в случае постоянных сдвигов — лишь линейная, хотя и с малыми, как правило, знаменателями геометрической прогрессии. Зачастую бывает достаточно сделать 2-3 итерации по формулам (4.24), чтобы получить заданную собственную пару с реально возможной точностью. Нужно только видеть разницу в цене реализации обратных итераций с постоянными и с переменными сдвигами: в первом случае при каждом k решаются линейные системы с одной и той же матрицей коэффициентов (как это было и при обратных итерациях (4.17) без сдвигов), во втором случае на разных шагах приходится решать совершенно различные системы.

В методе (4.24) также, как и в предыдущем, неясно, как подбирать начальный сдвиг $\sigma = \lambda_j^{(0)}$, за исключением случаев, когда решается частичная проблема заведомо в такой постановке, при которой требуется найти собственное число, ближайшее к заданному значению, и соответствующий ему собственный вектор.

Более определенной в этом смысле, к тому же более быстро сходящейся является следующая модификация метода (4.24) — ***обратные итерации с отношениями Рэлея***, применяемые для решения симметричных задач на собственные значения.

Ее основу составляет **RQI-*алгоритм****):

Шаг 0. Задать вектор $\mathbf{x}^{(0)}$ такой, что $\|\mathbf{x}^{(0)}\| = 1$.

Шаг 1. Для k=1, 2, ...:

1.1. Вычислить $\rho_{k-1} = (\mathbf{Ax}^{(k-1)}, \mathbf{x}^{(k-1)}) / (\mathbf{x}^{(k-1)}, \mathbf{x}^{(k-1)})$.

1.2. Найти $\mathbf{y}^{(k)}$ из уравнения $(\mathbf{A} - \rho_{k-1}\mathbf{E})\mathbf{y}^{(k)} = \mathbf{x}^{(k-1)}$.

1.3. Нормировать $\mathbf{y}^{(k)}$, т.е. положить $\mathbf{x}^{(k)} = \mathbf{y}^{(k)} / \|\mathbf{y}^{(k)}\|$.

1.4. Проверить ρ_{k-1}, $\mathbf{x}^{(k)}$ на сходимость. Перейти на шаг 1.1 или остановиться.

После «штатного» останова работы алгоритма**) при некотором $k = k_0$ в качестве собственной для данной матрицы **A** объявляется пара $\{\rho_{k_0-1}, \mathbf{x}^{(k_0)}\}$ или делается еще шаг 1.1 и берется $\{\rho_{k_0}, \mathbf{x}^{(k_0)}\}$.

---

*) RQI — *Rayleigh quotient iteration* (англ.).

**) Один из вариантов останова: $\|\mathbf{y}^{(k)}\| > C$, где $C > 0$ — большая константа [141].

Сдвиги на отношения Рэлея при наличии ортонормированного базиса из собственных векторов $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ обеспечивают асимптотически кубическую скорость сходимости ***последовательности Рэлея*** $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ к некоторому из векторов этого базиса [43, 141]. К какому именно, зависит от выбора начального вектора этой последовательности; беря различные линейно независимые векторы $\mathbf{x}^{(0)}$, можно получать разные собственные пары данной симметричной матрицы $\mathbf{A}$. При этом, правда, без дополнительных условий (типа (4.18) применительно к ρ_0 в роли σ) нельзя гарантировать, что найденное как предел последовательности $\rho_0, \rho_1, \rho_2, \ldots$ собственное число будет ближайшим к числу ρ_0.

Чтобы если не обосновать, то хотя бы осмыслить RQI-алгоритм, нужно вспомнить свойство 4.6, согласно которому, при выбранном векторе $\mathbf{x}^{(0)}$ вычисленное на первом шаге при $k=1$ отношение Рэлея $\rho_0 = (\mathbf{Ax}^{(0)}, \mathbf{x}^{(0)}) / (\mathbf{x}^{(0)}, \mathbf{x}^{(0)})$ можно считать некоторым приближением к собственному числу, связанному с заданным в R_n направлением $\mathbf{x}^{(0)}$. С этим начальным приближением к какому-то собственному числу λ_j далее выполняются обратные итерации с переменными сдвигами, как и в (4.24), только приближения к λ_j находятся не через отношения координат векторов $\mathbf{y}^{(k)}$ и $\mathbf{x}^{(k-1)}$, а через отношения Рэлея (как в методе скалярных произведений, см. замечание 4.2 в § 4.2), причем поскольку здесь ρ_k приближает собственное число данной матрицы $\mathbf{A}$, а не «сдвинутой», нет необходимости корректировать получаемое значение на величину смещения спектра, что имело место в формуле (4.23) и в последней из формул (4.24).

Замечание 4.7. RQI-алгоритм допускает использование любых векторных норм. Более естественно здесь применение евклидовой нормы; в таком случае, в силу $\|\mathbf{x}^{(k-1)}\|_2 = \sqrt{(\mathbf{x}^{(k-1)}, \mathbf{x}^{(k-1)})} = 1$, вычисление ρ_{k-1} можно производить по формуле

$$\rho_{k-1} = (\mathbf{Ax}^{(k-1)}, \mathbf{x}^{(k-1)}).$$

Замечание 4.8. Ясно, что применение переменных сдвигов в методе обратных итераций сильно ухудшает от шага к шагу обусловленность матриц решаемых там СЛАУ (они быстро приближаются к вырожденным). Однако, как показали проведенные Уилкинсоном [179] исследования, это не сказывается на достижимой точности получаемых таким методом результатов. Более того, Парлетт [141] аргументировано утверждает полезность плохой обусловленности (редкий случай!) матриц линейных систем в методах обратных итераций с хорошими сдвигами; объяснением

этому парадоксальному явлению служит сосредоточение ошибок округлений именно в направлении искомого собственного вектора, что только ускоряет доминирование нужной составляющей.

Пример 4.2. К симметричной матрице $\mathbf{A} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ из примера 4.1 применим RQI-алгоритм.

Примем за начальный вектор $\mathbf{x}^{(0)} = \frac{1}{\sqrt{5}}\begin{pmatrix} 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 0.6 \\ 0.8 \end{pmatrix}$ с нормой $\left\|\mathbf{x}^{(0)}\right\|_2 = \sqrt{\left(\mathbf{x}^{(0)},\ \mathbf{x}^{(0)}\right)} = 1$ и посмотрим, как поведет себя процесс обратных итераций с отношениями Рэлея (которые, в силу замечания 4.7, будем подсчитывать по упрощенной формуле). Следуя алгоритму, последовательно получаем:

$$\rho_0 = \left(\mathbf{A}\mathbf{x}^{(0)},\ \mathbf{x}^{(0)}\right) = \left(\begin{pmatrix} 0.4 \\ 1 \end{pmatrix}, \begin{pmatrix} 0.6 \\ 0.8 \end{pmatrix}\right) = 1.04;$$

$$(\mathbf{A} - \rho_0\mathbf{E})\mathbf{y}^{(1)} = \mathbf{x}^{(0)} \quad \Leftrightarrow \quad \begin{cases} 0.96\, y_1^{(1)} - y_2^{(1)} = 0.6, \\ -\, y_1^{(1)} + 0.96 y_2^{(1)} = 0.8 \end{cases} \quad \Leftrightarrow \quad \mathbf{y}^{(1)} = \begin{pmatrix} -17.551020 \\ -17.448980 \end{pmatrix},$$

$$\left\|\mathbf{y}^{(1)}\right\|_2 \approx 24.748843, \qquad \mathbf{x}^{(1)} = \frac{\mathbf{y}^{(1)}}{\left\|\mathbf{y}^{(1)}\right\|_2} \approx \begin{pmatrix} -0.709165 \\ -0.705043 \end{pmatrix};$$

далее при $k = 2$ аналогично имеем:

$$\rho_1 = \left(\mathbf{A}\mathbf{x}^{(1)},\ \mathbf{x}^{(1)}\right) \approx \left(\begin{pmatrix} -0.713288 \\ -0.700921 \end{pmatrix}, \begin{pmatrix} -0.709165 \\ -0.705043 \end{pmatrix}\right) \approx 1.000018;$$

$$(\mathbf{A} - \rho_1\mathbf{E})\mathbf{y}^{(2)} = \mathbf{x}^{(1)} \quad \Leftrightarrow \quad \mathbf{y}^{(2)} \approx \begin{pmatrix} 39283.201 \\ 39283.203 \end{pmatrix},$$

$$\left\|\mathbf{y}^{(2)}\right\|_2 \approx 55554.84, \qquad \mathbf{x}^{(2)} = \frac{\mathbf{y}^{(2)}}{\left\|\mathbf{y}^{(2)}\right\|_2} \approx \begin{pmatrix} 0.70710674 \\ 0.70710678 \end{pmatrix};$$

Вектор $\mathbf{x}^{(2)}$ и скаляр $\rho_2 = \left(\mathbf{A}\,\mathbf{x}^{(2)},\ \mathbf{x}^{(2)}\right) \approx 0.99999994$ с высокой точностью представляют младшую собственную пару $(\lambda_2,\ \mathbf{x}_2)$ данной матрицы $\mathbf{A}$.

Кроме предсказанной высокой скорости сходимости метода, обратим внимание на быстрый рост здесь величин $\left\|\mathbf{y}^{(k)}\right\|$, что, как ранее подмечено, можно положить в основу критерия окончания итерационного процесса. Использование одного естественного критерия $|\rho_k - \rho_{k-1}| < \varepsilon$ в некоторых случаях может подвести по причине зацикливания. Чтобы убедиться в этом, достаточно провести в условиях данного примера простейшие, буквально устные, вычисления, начинаемые с вектора $\mathbf{x}^{(0)} = (0;\ 1)^T$. Подчеркнем, что выбор начального вектора в данном методе играет весьма существенную роль.

4.4. МЕТОД ВРАЩЕНИЙ ЯКОБИ РЕШЕНИЯ СИММЕТРИЧНОЙ ПОЛНОЙ ПРОБЛЕМЫ СОБСТВЕННЫХ ЗНАЧЕНИЙ

Дальнейшее изучение методов решения алгебраических проблем собственных значений существенно опирается на матричное ***преобразование подобия***. Напомним, что *подобными называются матрицы* $\mathbf{A}$ *и* $\mathbf{B} = \mathbf{C}^{-1}\mathbf{A}\mathbf{C}$, *где* $\mathbf{C}$ *— произвольная невырожденная матрица.*

Пополним приведенный в § 4.1 набор простейших свойств собственных пар матриц еще двумя свойствами.

Свойство 4.7. *Пусть* $\{\lambda,\ \mathbf{x}\}$ *— собственная пара матрицы* $\mathbf{B} = \mathbf{C}^{-1}\mathbf{A}\mathbf{C}$. *Тогда* $\{\lambda,\ \mathbf{C}\mathbf{x}\}$ *— собственная пара матрицы* $\mathbf{A}$.

Чтобы убедиться в справедливости этого свойства, достаточно подставить выражение $\mathbf{B} = \mathbf{C}^{-1}\mathbf{A}\mathbf{C}$ в верное для пары $\{\lambda,\ \mathbf{x}\}$ равенство $\mathbf{B}\mathbf{x} = \lambda\mathbf{x}$: имеем $\mathbf{C}^{-1}\mathbf{A}\mathbf{C}\mathbf{x} = \lambda\mathbf{x}$, откуда после умножения слева на матрицу $\mathbf{C}$ получаем равенство $\mathbf{A}\mathbf{C}\mathbf{x} = \lambda\mathbf{C}\mathbf{x}$, означающее истинность утверждения.

Как видим, *преобразование подобия сохраняет неизменным спектр любой матрицы.*

Свойство 4.8. *Пусть* $\mathbf{A}$ *—* $n \times n$*–матрица простой структуры* (см. § 4.2), *а матрицы* $\mathbf{\Lambda} = \mathrm{diag}(\lambda_i)$ *и* $\mathbf{X} = (\mathbf{x}_1;\ \mathbf{x}_2;\ \ldots;\ \mathbf{x}_n)$ *образованы из ее собственных чисел и собственных векторов соответственно. Тогда справедливо равенство* $\mathbf{\Lambda} = \mathbf{X}^{-1}\mathbf{A}\mathbf{X}$.

Действительно, то, что $\{\lambda_i,\ \mathbf{x}_i\}$ являются собственными парами матрицы $\mathbf{A}$, означает, что

$$\mathbf{A}\mathbf{x}_i = \lambda_i\mathbf{x}_i \quad \forall i \in \{1,\ 2,\ \ldots,\ n\}.$$

Эти n равенств могут быть записаны в виде одного матричного равенства

$$\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{\Lambda}. \tag{4.25}$$

В силу простой структуры $\mathbf{A}$, все ее собственные векторы, т.е. столбцы матрицы $\mathbf{X}$, линейно независимы, поэтому матрица $\mathbf{X}$ обратима. Умножив равенство (4.25) слева на матрицу $\mathbf{X}^{-1}$, получим нужное представление $\mathbf{\Lambda} = \mathbf{X}^{-1}\mathbf{A}\mathbf{X}$.

Так как для диагональной матрицы $\mathbf{\Lambda}$, образованной из собственных чисел, собственными векторами могут служить

6 Б-223

единичные векторы исходного базиса (действительно, $\mathbf{A}\mathbf{e}_i = \lambda_i \mathbf{e}_i \quad \forall i \in \{1, 2, \ldots, n\}$), то применяя к последнему случаю свойство 4.7 с $\mathbf{C}=\mathbf{X}$ и с $\mathbf{x} = \mathbf{e}_i$ (т.е. с $\mathbf{C}\mathbf{x} = \mathbf{X}\mathbf{e}_i = \mathbf{x}_i$), приходим к другой формулировке свойства 4.8:

> *если* $\{\lambda_i, \mathbf{e}_i\}$ *является собственной парой матрицы* $\mathbf{\Lambda} = \mathrm{diag}(\lambda_i) = \mathbf{X}^{-1}\mathbf{A}\mathbf{X}$, *то* $\{\lambda_i, \mathbf{x}_i\}$ *есть собственная пара матрицы* $\mathbf{A}$ (обозначения те же, что и в свойстве 4.8).

Далее (в пределах этого параграфа) будем рассматривать только симметричные вещественные матрицы. Пользуясь известным фактом о наличии у таких матриц полной ортонормированной системы собственных векторов, т.е. тем, что заявленная выше матрица $\mathbf{X}$ из собственных векторов в этом случае является ортогональной ($\mathbf{X}^{-1} = \mathbf{X}^T$), запишем как следствие свойства 4.8 равенство

$$\mathbf{\Lambda} = \mathbf{X}^T \mathbf{A}\mathbf{X}. \qquad (4.26)$$

Значит, для всякой симметричной матрицы $\mathbf{A}$ найдется диагональная матрица $\mathbf{\Lambda}$, ей ортогонально подобная. Вопрос теперь состоит в том, как реализовать хотя бы приближенно равенство (4.26), которое позволило бы найти сразу все собственные числа матрицы $\mathbf{A}$ (элементы диагонали матрицы $\mathbf{\Lambda}$) и все соответствующие им собственные векторы (столбцы матрицы $\mathbf{X}$)? Один из возможных ответов на этот вопрос состоит в применении к $\mathbf{A}$ последовательности однотипных преобразований, сохраняющих спектр и приводящих в пределе данную матрицу к диагональному виду.

Для этих целей будем использовать преобразования с помощью так называемой ***матрицы плоских вращений***

$$\mathbf{T}_{ij} = \begin{array}{cc} \begin{array}{ccccc} & i & & j & \end{array} & \\ \begin{pmatrix} 1 & \vdots & & \vdots & 0 \\ \cdots & c & \cdots & -s & \cdots \\ & \vdots & & \vdots & \\ \cdots & s & \cdots & c & \cdots \\ 0 & \vdots & & \vdots & 1 \end{pmatrix} & \begin{matrix} \\ i \\ \\ j \\ \\ \end{matrix} \end{array} . \qquad (4.27)$$

Она получается из единичной матрицы заменой двух единиц и двух нулей на пересечениях i-х и j-х строк и столбцов числами c и $\pm s$, как показано в (4.27), такими, что

$$c^2 + s^2 = 1. \qquad (4.28)$$

Условие нормировки (4.28) позволяет интерпретировать числа c и s как косинус и синус некоторого угла α, и, так как умножение любой матрицы на матрицу $\mathbf{T}_{ij}$ изменяет у нее только две строки и два столбца по формулам поворота на угол α в плоскости, определяемой выбранной парой индексов i и j, то это полностью оправдывает название матрицы $\mathbf{T}_{ij}$.

Матрица $\mathbf{T}_{ij}$ ортогональна при любых $i, j \in \{1, 2, \ldots, n\}$ (проверьте!), и значит, матрица

$$\mathbf{B} = \mathbf{T}_{ij}^T \mathbf{A}\mathbf{T}_{ij} \qquad (4.29)$$

подобна $\mathbf{A}$, т.е. имеет тот же набор собственных чисел, что и матрица $\mathbf{A}$.

Классический итерационный метод вращений, предложенный Якоби (1846 г.), предполагает построение последовательности матриц

$$\mathbf{B}_0 (:= \mathbf{A}), \mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_k, \ldots$$

с помощью преобразований типа (4.29)

$$\mathbf{B}_k = \mathbf{T}_{ij}^T \mathbf{B}_{k-1} \mathbf{T}_{ij} \qquad (4.30)$$

такой, что на k-м шаге *обнуляется максимальный по модулю элемент* матрицы $\mathbf{B}_{k-1}$ предыдущего шага (а значит, и симметричный ему элемент). Эта стратегия определяет способ фиксирования пары индексов i, j, задающих позиции $(i,i), (j,j), (i,j), (j,i)$ «существенных» элементов в матрице вращения $\mathbf{T}_{ij}$, и угол поворота α, конкретизирующий значения этих элементов $c = \cos\alpha$ и $\pm s = \pm\sin\alpha$. На каждом шаге таких преобразований пересчитываются только две строки (или два столбца, что неважно в силу симметрии) матрицы предыдущего шага. Хотя, к сожалению, нельзя рассчитывать, что таким путем за конечное число шагов можно точно найти диагональную матрицу $\mathbf{\Lambda}$, ибо полученные на некотором этапе преобразований нулевые элементы на следующем этапе станут, вообще говоря, ненулевыми, но нужное предельное поведение

$$\mathbf{B}_k \xrightarrow[k\to\infty]{} \mathbf{\Lambda},$$

как будет показано ниже, есть.

Определив идею метода вращений, рассмотрим теперь его несколько подробней.

Пусть $\mathbf{A} = (a_{ml})_{m,\,l=1}^n$ — исходная симметричная матрица, а $\mathbf{B} = (b_{ml})_{m,\,l=1}^n$ — матрица, получающаяся после одного шага

преобразований по формуле (4.29). Обозначим соответственно через $\widetilde{\mathbf{A}}$ и $\widetilde{\mathbf{B}}$ двумерные подматрицы этих матриц*), определяемые фиксированием позиции (i, j) некоторого элемента a_{ij} матрицы $\mathbf{A}$:

$$\widetilde{\mathbf{A}} = \begin{pmatrix} a_{ii} & a_{ij} \\ a_{ij} & a_{jj} \end{pmatrix}, \quad \widetilde{\mathbf{B}} = \begin{pmatrix} b_{ii} & b_{ij} \\ b_{ji} & b_{jj} \end{pmatrix},$$

а через $\widetilde{\mathbf{T}}$ — такую же подматрицу матрицы $\mathbf{T}_{ij}$:

$$\widetilde{\mathbf{T}} = \begin{pmatrix} c & -s \\ s & c \end{pmatrix}.$$

Очевидно, что равенство (4.29), записанное для матриц $\mathbf{A}, \mathbf{B}, \mathbf{T}_{ij}$, будет верным и для их подматриц $\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}}, \widetilde{\mathbf{T}}$. Пользуясь этим, подсчитаем элементы матрицы $\widetilde{\mathbf{B}}$, выполняя умножение в правой части двумерного аналога (4.29):

$$\widetilde{\mathbf{B}} = \widetilde{\mathbf{T}}^T \widetilde{\mathbf{A}} \widetilde{\mathbf{T}} = \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} ca_{ii} + sa_{ij} & ca_{ij} - sa_{ii} \\ ca_{ij} + sa_{jj} & ca_{jj} - sa_{ij} \end{pmatrix} =$$

$$= \begin{pmatrix} c^2 a_{ii} + 2csa_{ij} + s^2 a_{jj} & c^2 a_{ij} - csa_{ii} + csa_{jj} - s^2 a_{ij} \\ c^2 a_{ij} - csa_{ii} + csa_{jj} - s^2 a_{ij} & c^2 a_{jj} - 2csa_{ij} + s^2 a_{ii} \end{pmatrix}.$$

Отсюда видим, что $b_{ij} = b_{ji} = 0$, если

$$(c^2 - s^2) a_{ij} - cs(a_{ii} - a_{jj}) = 0,$$

т.е. если

$$\frac{cs}{c^2 - s^2} = \frac{a_{ij}}{a_{ii} - a_{jj}}.$$

Учитывая тригонометрическую интерпретацию чисел $c = \cos\alpha$ и $s = \sin\alpha$, в соответствии с чем можно считать

$$cs = \frac{\sin 2\alpha}{2}, \quad c^2 - s^2 = \cos 2\alpha,$$

*) В $\widetilde{\mathbf{A}}$ элемент a_{ji} сразу заменяем равным ему элементом a_{ij}, с равенством же $b_{ji} = b_{ij}$ в $\widetilde{\mathbf{B}}$ пока не торопимся.

приходим к выводу, что матрица $\mathbf{B}$ будет иметь нулевые внедиагональные элементы $b_{ij} = b_{ji}$, если использовать преобразование плоского вращения по формуле (4.29) на угол α такой, что

$$\operatorname{tg} 2\alpha = \frac{2a_{ij}}{a_{ii} - a_{jj}}$$

(для определенности считают $\alpha \in \left(-\frac{\pi}{4}, \frac{\pi}{4}\right]$).

Ясно, что нет необходимости находить непосредственно угол α, поскольку нужные для выполнения преобразований числа c и s можно получить через значение $\operatorname{tg} 2\alpha$ по формулам тригонометрии. При этом сразу отметим, что наибольшие требования к точности в описываемом методе предъявляются именно на стадии вычисления c и s, так как здесь возможны наибольшие потери точности, а искажение c и s нарушает ортогональность матриц $\mathbf{T}$, что ведет к неустранимым погрешностям (метод вращений, итерационный по форме, не является итерационным по существу: ему не присуща самоисправляемость методов последовательных приближений).

Проделав соответствующие элементарные выкладки и возвратившись к n-мерному случаю, запишем теперь совокупность формул, определяющую **один шаг метода вращений Якоби**. Для того чтобы не перегружать эти формулы лишними индексами, будем считать что преобразуется матрица $\mathbf{A}$ в матрицу $\mathbf{B}$ согласно (4.29), хотя на самом деле на k-м шаге должно применяться преобразование (4.30) к матрице $\mathbf{B}_{k-1} = \left(b_{ml}^{(k-1)}\right)$ с результатом $\mathbf{B}_k = \left(b_{ml}^{(k)}\right)$.

Итак, пусть a_{ij} — ***ключевой элемент*** преобразуемой матрицы $\mathbf{A}$. Матрица $\mathbf{B}$, подобная $\mathbf{A}$, формируется следующим образом:

1. Вычисляют $p := 2a_{ij}$, $q := a_{ii} - a_{jj}$, $d := \sqrt{p^2 + q^2}$.

2. Если $q \neq 0$, то $r := |q|/(2d)$, $c := \sqrt{0.5 + r}$, $s := \sqrt{0.5 - r} \cdot \operatorname{sign}(pq)$ (если $|p| << |q|$, то лучше $s := |p| \cdot \operatorname{sign}(pq)/(2cd)$),
если же $q = 0$, то $c = s := \sqrt{2}/2$.

3. Вычисляют новые диагональные элементы:

$$b_{ii} := c^2 a_{ii} + s^2 a_{jj} + 2csa_{ij},$$

$$b_{jj} := s^2 a_{ii} + c^2 a_{jj} - 2csa_{ij}$$

4. Полагают $b_{ij} = b_{ji} := 0$ (или для контроля вычисляют

$$b_{ij} = b_{ji} := \left(c^2 - s^2\right)a_{ij} + cs\left(a_{jj} - a_{ii}\right));$$

5. При $m = 1, 2, \ldots, n$ таких, что $m \neq i$, $m \neq j$, вычисляют изменяющиеся внедиагональные элементы:

$$\begin{aligned} b_{im} &= b_{mi} := ca_{mi} + sa_{mj}, \\ b_{jm} &= b_{mj} := -sa_{mi} + ca_{mj}. \end{aligned} \tag{4.31}$$

6. Для всех остальных пар индексов m, l принимают $b_{ml} := a_{ml}$.

Конечно, в реальных вычислениях, если это считать основой алгоритма, не все записанное здесь следует выполнять, а именно, не нужно делать последних переприсвоений, а также должна учитываться симметрия получающейся матрицы $\mathbf{B}$.

Убедимся теперь, что *если в качестве ключевого или, в иной терминологии* [141], ***обреченного элемента*** *на каждом шаге преобразований подобия по указанным формулам брать максимальный по модулю элемент преобразуемой матрицы, то в пределе получится диагональная матрица.*

Для доказательства этого, т.е. доказательства сходимости последовательности $(\mathbf{B}_k)$ к $\mathbf{\Lambda}$, используем (см. приложение 1) норму Фробениуса

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i,j=1}^{n} a_{ij}^2}.$$

Проследим за поведением норм матриц (точнее, квадратов этих норм), получающихся из матриц $\mathbf{B}_k$ заменой диагональных элементов нулями. Такие матрицы, определяемые внедиагональными элементами матриц $\mathbf{B}_k$, будем обозначать $\mathbf{V}_{\mathbf{B}_k}$. При этом опять для упрощения записей будем пока рассматривать переход от $\mathbf{A}$ к $\mathbf{B}$.

Найдем выражение суммы квадратов внедиагональных элементов матрицы

$$\mathbf{B} = \begin{pmatrix} a_{11} & \ldots & b_{1i} & \ldots & b_{1j} & \ldots & a_{1n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ b_{i1} & \ldots & b_{ii} & \ldots & 0 & \ldots & b_{in} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ b_{j1} & \ldots & 0 & \ldots & b_{jj} & \ldots & b_{jn} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & \ldots & b_{ni} & \ldots & b_{nj} & \ldots & a_{nn} \end{pmatrix},$$

принадлежащих изменяющимся по сравнению с $\mathbf{A}$ строке i и столбцу j, через элементы матрицы $\mathbf{A}$. Имеем (с учетом (4.31), (4.28) и условия обнуления элементов $b_{ij} = b_{ji}$):

$$\begin{aligned} \sum_{m\neq i} b_{im}^2 + \sum_{m\neq j} b_{mj}^2 = &\sum_{m\neq i,\, j}\left(c^2 a_{mi}^2 + 2csa_{mi}a_{mj} + s^2 a_{mj}^2\right) + \\ &+ b_{ij}^2 + \sum_{m\neq j,\, i}\left(s^2 a_{mi}^2 - 2csa_{mi}a_{mj} + c^2 a_{mj}^2\right) + b_{ji}^2 = \\ &= \sum_{m\neq i,\, j}\left[\left(c^2 + s^2\right)a_{mi}^2 + \left(c^2 + s^2\right)a_{mj}^2\right] = \sum_{m\neq i,\, j}\left(a_{mi}^2 + a_{mj}^2\right). \end{aligned}$$

Аналогичные суммы квадратов j-й строки и i-го столбца, в силу симметрии, дадут такое же выражение. Это означает, что если полученное равенство удвоить и дополнить левую и правую части суммой квадратов всех остальных внедиагональных элементов матрицы $\mathbf{A}$ (служащих соответствующими элементами и матрицы $\mathbf{B}$), то в левой части будет стоять сумма квадратов всех внедиагональных элементов матрицы $\mathbf{B}$, а в правой части — сумма квадратов всех внедиагональных элементов матрицы $\mathbf{A}$, кроме a_{ji}^2 и a_{ij}^2. Следовательно, справедливо равенство

$$\|\mathbf{V_B}\|_F^2 = \|\mathbf{V_A}\|_F^2 - 2a_{ij}^2, \tag{4.32}$$

говорящее об *убывании сумм квадратов внедиагональных элементов в рассматриваемом процессе преобразований подобия.*

Пусть $\left|a_{ij}\right| = \max\limits_{m\neq l}\{|a_{ml}|\}$. Тогда можно считать, что $a_{ij}^2 = \max\limits_{m\neq l}\{a_{ml}^2\}$ не меньше, чем среднее значение множества из

$n^2 - n$ квадратов всех внедиагональных элементов n-мерной матрицы $\mathbf{A}$, т.е. величины

$$\frac{1}{n(n-1)}\sum_{m\neq l} a_{ml}^2 = \frac{1}{n(n-1)}\|\mathbf{V_A}\|_F^2.$$

Подставляя полученную оценку $a_{ij}^2 \geq \frac{1}{n(n-1)}\|\mathbf{V_A}\|_F^2$ в равенство (4.32), приходим к неравенству

$$\|\mathbf{V_B}\|_F^2 \leq \left(1-\frac{2}{n(n-1)}\right)\|\mathbf{V_A}\|_F^2.$$

На основании этого для последовательности матриц $\mathbf{B}_k$, представляемых в виде

$$\mathbf{B}_k = \operatorname{diag}\left(b_{ii}^{(k)}\right) + \mathbf{V}_{\mathbf{B}_k},$$

можно записать:

$$\|\mathbf{V}_{\mathbf{B}_k}\|_F^2 \leq \left(1-\frac{2}{n(n-1)}\right)\|\mathbf{V}_{\mathbf{B}_{k-1}}\|_F^2 \leq \ldots \leq$$

$$\leq \left(1-\frac{2}{n(n-1)}\right)^k \|\mathbf{V_A}\|_F^2 \xrightarrow[k\to\infty]{} 0.$$

Значит, при указанном способе выбора ключевого элемента последовательность подобных матриц $\mathbf{B}_k$ сходится к диагональной матрице $\mathbf{\Lambda}$ из собственных значений, по крайней мере, со скоростью геометрической прогрессии.

Описанный выше классический вариант метода вращений Якоби, как показывают более тонкие оценки, на самом деле имеет асимптотически квадратичную скорость сходимости [42, 141, 179]. Однако при больших размерностях n его реализация наталкивается на существенные потери машинных ресурсов, связанные с поиском наибольшего по модулю ключевого элемента. Поэтому чаще применяется более медленно, но все-таки тоже асимптотически квадратично сходящийся ***циклический метод Якоби с барьерами***. Стратегия выбора ключевого элемента здесь такова: устраивается циклический перебор всех над- или поддиагональных элементов матрицы $\mathbf{A}$ (точнее, $\mathbf{B}_{k-1}$) для их использования в роли обреченного, но при этом пропускаются элементы, абсолютные величины которых меньше некоторого заданного

положительного числа — барьера. Этот барьер может быть переменным (уменьшающимся по какому-либо осмысленному принципу).

Следует отметить успешную работу таких итерационных процессов и в случаях, когда исходная матрица имеет кратные и, что хуже, близкие собственные числа.

В завершение, остается вспомнить, что в соответствии со свойствами 4.7, 4.8 и проведенными рассуждениями за собственные векторы $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ матрицы $\mathbf{A}$ (имеющие единичную евклидову норму) могут быть приближенно приняты столбцы результирующей матрицы, получающейся справа от $\mathbf{A}$ в цепочке преобразований подобия

$$\mathbf{B}_1 = \mathbf{T}_{i_0 j_0}^T \mathbf{A} \mathbf{T}_{i_0 j_0},$$

$$\mathbf{B}_2 = \mathbf{T}_{i_1 j_1}^T \mathbf{B}_1 \mathbf{T}_{i_1 j_1} = \mathbf{T}_{i_1 j_1}^T \mathbf{T}_{i_0 j_0}^T \mathbf{A} \mathbf{T}_{i_0 j_0} \mathbf{T}_{i_1 j_1},$$

. . . .

$$\mathbf{B}_k = \mathbf{T}_{i_{k-1} j_{k-1}}^T \ldots \mathbf{T}_{i_0 j_0}^T \mathbf{A} \mathbf{T}_{i_0 j_0} \ldots \mathbf{T}_{i_{k-1} j_{k-1}},$$

т.е. матрицы

$$\mathbf{T}_k := \mathbf{T}_{i_0 j_0} \cdot \mathbf{T}_{i_1 j_1} \cdot \ldots \cdot \mathbf{T}_{i_{k-1} j_{k-1}}$$

при некотором $k = K$. Разумеется, в реальном вычислительном процессе нужно обойтись без матричных умножений.

Сигналом для окончания процесса вращений Якоби может служить, например, достаточная малость $\left\|\mathbf{V}_{\mathbf{B}_k}\right\|_F$ или $\left|b_{ij}^{(k)}\right|$.

Пример 4.3. Методом вращений Якоби решим задачу нахождения всех собственных пар матрицы $\mathbf{A} = \begin{pmatrix} 1 & 1 & 3 \\ 1 & 5 & 1 \\ 3 & 1 & 1 \end{pmatrix}$.

К данной симметричной (отрицательно определенной) матрице $\mathbf{A}$ будем поэтапно применять записанный выше основной фрагмент алгоритма, переводящий ее в матрицу $\mathbf{B}$, являющуюся подобной $\mathbf{A}$ и имеющую меньшую сумму квадратов внедиагональных элементов. При этом условимся, что ключевой элемент a_{ij} будем брать в поддиагональной части матрицы $\mathbf{A}$ (можно и наоборот, ничего от этого принципиально не изменится) и что все вычисления будем проводить точно (т.е. рассматривается идеальный процесс).

Этап 1*-й*. Выбираем ключевой элемент $a_{31} = 3$ (максимальный в условленной части матрицы), следовательно, фиксируем индексы $i = 3$,

$j = 1$. Вычисляем

$$p = 2a_{31} = 6, \quad q = a_{33} - a_{11} = 1 - 1 = 0 \;\Rightarrow\; c = s = \frac{1}{\sqrt{2}}.$$

Далее находим элементы новой матрицы

$$b_{33} = c^2 a_{33} + s^2 a_{11} + 2csa_{31} = \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 1 + 2\cdot\frac{1}{2}\cdot 3 = 4,$$

$$b_{11} = s^2 a_{33} + c^2 a_{11} - 2csa_{31} = \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 1 - 2\cdot\frac{1}{2}\cdot 3 = -2$$

и при $m = 2$

$$b_{32} = b_{23} = ca_{23} + sa_{21} = \frac{1}{\sqrt{2}}(1+1) = \sqrt{2},$$

$$b_{12} = b_{21} = -sa_{23} + ca_{21} = \frac{1}{\sqrt{2}}(-1+1) = 0.$$

Следовательно,

$$\mathbf{A} \infty \mathbf{B} = \begin{pmatrix} -2 & 0 & 0 \\ 0 & 5 & \sqrt{2} \\ 0 & \sqrt{2} & 4 \end{pmatrix}.$$

Этап 2-*й*. Полученную на предыдущем этапе матрицу $\mathbf{B}$ считаем матрицей $\mathbf{A}$, т.е. будем считать по тем же формулам, положив $\mathbf{A} := \mathbf{B}$. Ключевой элемент здесь $a_{32} = \sqrt{2}$, т.е. $i = 3$, $j = 2$. Согласно алгоритму, имеем:

$$p = 2a_{32} = 2\sqrt{2}, \quad q = a_{33} - a_{22} = 4 - 5 = -1 \;\text{(замечаем, что } pq < 0\text{)},$$

$$d = \sqrt{p^2 + q^2} = \sqrt{8+1} = 3, \quad r = \frac{|q|}{2d} = \frac{1}{6},$$

$$c = \sqrt{0.5 + r} = \sqrt{\frac{1}{2} + \frac{1}{6}} = \frac{2}{\sqrt{6}}, \quad s = \sqrt{0.5 - r}\cdot sign(pq) = -\sqrt{\frac{1}{2} - \frac{1}{6}} = -\frac{\sqrt{2}}{\sqrt{6}}.$$

Зная числа c и s, определяющие преобразование вращения, вычисляем

$$b_{33} = c^2 a_{33} + s^2 a_{22} + 2csa_{32} = \frac{4}{6}\cdot 4 + \frac{2}{6}\cdot 5 + 2\cdot\frac{2}{\sqrt{6}}\cdot\left(-\frac{\sqrt{2}}{\sqrt{6}}\right)\cdot\sqrt{2} = 3,$$

$$b_{22} = s^2 a_{33} + c^2 a_{22} - 2csa_{32} = \frac{2}{6}\cdot 4 + \frac{4}{6}\cdot 5 + \frac{8}{6} = 6$$

и далее при $m = 1$

$$b_{31} = b_{13} = ca_{13} + sa_{12} = \frac{2}{\sqrt{6}}\cdot 0 + \left(-\frac{\sqrt{2}}{\sqrt{6}}\right)\cdot 0 = 0,$$

$$b_{21} = b_{12} = -sa_{13} + ca_{12} = \frac{\sqrt{2}}{\sqrt{6}}\cdot 0 + \frac{2}{\sqrt{6}}\cdot 0 = 0.$$

Таким образом, уже после второго этапа (это не норма, а, будем считать, везение) получена диагональная матрица

$$\mathbf{B} = \begin{pmatrix} -2 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 3 \end{pmatrix},$$

подобная исходной матрице $\mathbf{A}$, в силу чего можно утверждать, что собственными значениями матрицы $\mathbf{A}$ являются числа

$$\lambda_1 = -2, \quad \lambda_2 = 6, \quad \lambda_3 = 3.$$

Чтобы найти отвечающие им собственные векторы, выпишем матрицы плоских вращений первого и второго этапов (в соответствии с их структурой (4.27))

$$\mathbf{T}_{ij}^{(1)} = \mathbf{T}_{31} = \begin{pmatrix} c^{(1)} & 0 & s^{(1)} \\ 0 & 1 & 0 \\ -s^{(1)} & 0 & c^{(1)} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{pmatrix},$$

$$\mathbf{T}_{ij}^{(2)} = \mathbf{T}_{32} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c^{(2)} & s^{(2)} \\ 0 & -s^{(2)} & c^{(2)} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{2}{\sqrt{6}} & -\frac{\sqrt{2}}{\sqrt{6}} \\ 0 & \frac{\sqrt{2}}{\sqrt{6}} & \frac{2}{\sqrt{6}} \end{pmatrix}.$$

Их произведение, т.е. матрица

$$\mathbf{T} = \mathbf{T}_{31}\cdot\mathbf{T}_{32} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{\sqrt{2}}{\sqrt{6}} \\ 0 & \frac{2}{\sqrt{6}} & -\frac{\sqrt{2}}{\sqrt{6}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{\sqrt{2}}{\sqrt{6}} \end{pmatrix},$$

согласно последним теоретическим выкладкам этого параграфа, имеет своими столбцами собственные векторы матрицы $\mathbf{A}$. Легко проверить по определению, что ее столбцы в естественном порядке, т.е. векторы

$$\mathbf{x}_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \quad \mathbf{x}_2 = \frac{1}{\sqrt{6}}\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \quad \mathbf{x}_3 = \frac{1}{\sqrt{3}}\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$$

образуют собственные пары с числами $\lambda_1 = -2$, $\lambda_2 = 6$ и $\lambda_3 = 3$ соответственно.

4.5. ПОНЯТИЕ ОБ LU-АЛГОРИТМЕ ДЛЯ НЕСИММЕТРИЧНЫХ ЗАДАЧ

Чаще (по крайней мере, в несимметричном случае) алгоритмы приближенного решения полных проблем собственных значений основываются на приведении данных матриц к подобным им матрицам не диагонального, а треугольного вида. Наиболее простой из таких алгоритмов вычисления собственных чисел опирается на хорошо известное LU-разложение матриц (см. гл. 2).

Пусть данная $n \times n$-матрица $\mathbf{A}$ представлена в виде $\mathbf{A} = \mathbf{LU}$, где $\mathbf{L}$ и $\mathbf{U}$ — соответственно нижняя и верхняя треугольные матрицы.

Обозначим[*] $\mathbf{A}_1 := \mathbf{UL}$, тогда $\mathbf{U} = \mathbf{A}_1\mathbf{L}^{-1}$. Подставив это выражение матрицы $\mathbf{U}$ в равенство $\mathbf{A} = \mathbf{LU}$, получаем новое представление $\mathbf{A}$:

$$\mathbf{A} = \mathbf{LA}_1\mathbf{L}^{-1}, \tag{4.33}$$

которое говорит о подобии матриц $\mathbf{A}$ и $\mathbf{A}_1$, т.е. о равенстве их собственных чисел $\lambda_{\mathbf{A}}$ и $\lambda_{\mathbf{A}_1}$.

Если матрица $\mathbf{A}_1$ может быть, как и $\mathbf{A}$, представлена в виде произведения нижней $\mathbf{L}_1$ и верхней $\mathbf{U}_1$ треугольных, т.е. $\mathbf{A}_1 = \mathbf{L}_1\mathbf{U}_1$, то, положив $\mathbf{A}_2 := \mathbf{U}_1\mathbf{L}_1$ и выразив отсюда $\mathbf{U}_1 = \mathbf{A}_2\mathbf{L}_1^{-1}$, аналогично предыдущему получаем

$$\mathbf{A}_1 = \mathbf{L}_1\mathbf{A}_2\mathbf{L}_1^{-1}. \tag{4.34}$$

Следовательно, $\mathbf{A}_1$ подобна $\mathbf{A}_2$ и ,значит, $\lambda_{\mathbf{A}_1} = \lambda_{\mathbf{A}_2}$.

Суперпозиция этих двух преобразований, т.е. подстановка (4.34) в (4.33), дает выражение $\mathbf{A}$ через $\mathbf{A}_2$:

$$\mathbf{A} = \mathbf{LL}_1\mathbf{A}_2\mathbf{L}_1^{-1}\mathbf{L}^{-1} = \mathbf{LL}_1\mathbf{A}_2(\mathbf{LL}_1)^{-1},$$

непосредственно утверждающее равенство собственных чисел $\lambda_{\mathbf{A}}$ и $\lambda_{\mathbf{A}_2}$.

Такой процесс построения теоретически бесконечной последовательности подобных матриц и составляет основу ***LU***- (иначе, ***LR***-) ***алгоритма***[**]. Он определяется фактически двумя

---

[*] Напомним, что произведение матриц, вообще говоря, некоммутативно.

[**] Алгоритм предложен Рутисхаузером (1958 г.).

формулами:

$$\mathbf{A}_k = \mathbf{L}_k\mathbf{U}_k, \quad \mathbf{A}_{k+1} = \mathbf{U}_k\mathbf{L}_k, \tag{4.35}$$

где $\mathbf{A}_0 := \mathbf{A}$, $k = 0, 1, 2, \ldots$, причем первая из этих формул означает процедуру треугольной факторизации матрицы $\mathbf{A}_k$ на k-м шаге, а вторая — простое умножение верхней треугольной матрицы на нижнюю.

Доказано [138, 179], что при ряде ограничений на данную матрицу $\mathbf{A}$ (простейшим из которых является, в частности, требование, чтобы все ее собственные числа были различны по модулю) итерационный процесс (4.35) осуществим, и формируемая им последовательность $(\mathbf{A}_k)$ сходится к треугольной матрице вида

$$\begin{pmatrix} \lambda_1 & * & * & \ldots & * \\ 0 & \lambda_2 & * & \ldots & * \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & \lambda_n \end{pmatrix} \quad \text{или вида} \quad \begin{pmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ * & \lambda_2 & 0 & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ * & * & * & \ldots & \lambda_n \end{pmatrix}$$

в зависимости от того, фиксируется единичная диагональ при LU-факторизации у матрицы $\mathbf{L}$ или у $\mathbf{U}$ соответственно (см. § 2.3). К сожалению, эти ограничения трудно назвать конструктивными, и реализующие LU-алгоритм программы больше опираются на эмпирику. Осуществимости, устойчивости и ускорения сходимости процесса (4.35) обычно добиваются (если это в принципе возможно) путем подходящих сдвигов матриц и перестановок их элементов; соответствующие исследования и рекомендации по этому поводу можно найти в [179].

Пример 4.4. Рассмотрим, как ведет себя LU-алгоритм (4.35), примененный к нахождению собственных чисел матрицы $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 6 & 1 \end{pmatrix}$.

Выполнив LU-разложение[*], получим

$$\mathbf{A}_0 := \mathbf{A} = \mathbf{L}_0\mathbf{U}_0 = \begin{pmatrix} 2 & 0 \\ 6 & -2 \end{pmatrix}\begin{pmatrix} 1 & 0.5 \\ 0 & 1 \end{pmatrix}.$$

---

[*] По формулам из § 2.3 $l_{ij} = a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj}$ $(i \ge j)$, $u_{ij} = \frac{1}{l_{ii}}\left(a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}\right)$ $(i < j)$, используемым поочередно.

Перемножая матрицы $\mathbf{L}_0$ и $\mathbf{U}_0$ в обратном порядке, строим матрицу

$$\mathbf{A}_1 := \mathbf{U}_0\mathbf{L}_0 = \begin{pmatrix} 5 & -1 \\ 6 & -2 \end{pmatrix}.$$

Факторизуя эту матрицу аналогично предыдущему, имеем

$$\mathbf{A}_1 = \mathbf{L}_1\mathbf{U}_1 = \begin{pmatrix} 5 & 0 \\ 6 & -0.8 \end{pmatrix}\begin{pmatrix} 1 & -0.2 \\ 0 & 1 \end{pmatrix},$$

откуда

$$\mathbf{A}_2 := \mathbf{U}_1\mathbf{L}_1 = \begin{pmatrix} 3.8 & 0.16 \\ 6 & -0.8 \end{pmatrix}.$$

Следующий шаг дает

$$\mathbf{A}_2 = \mathbf{L}_2\mathbf{U}_2 = \begin{pmatrix} 3.8 & 0 \\ 6 & -1.0526... \end{pmatrix}\begin{pmatrix} 1 & 0.0421... \\ 0 & 1 \end{pmatrix},$$

$$\mathbf{A}_3 := \mathbf{U}_2\mathbf{L}_2 = \begin{pmatrix} 4.0526... & 0.0443... \\ 6 & -1.0526... \end{pmatrix}.$$

Как видим, диагональные элементы матрицы $\mathbf{A}_2$ отличаются от точных значений собственных чисел $\lambda_1 = 4$, $\lambda_2 = -1$ на 0.2, а матрица $\mathbf{A}_3$ позволяет указать значения λ_1 и λ_2 с погрешностью ≈ 0.05.

Если в этом же примере фиксировать единичную диагональ у матриц $\mathbf{L}_k$, то процесс (4.35) будет развиваться следующим образом*):

$$\mathbf{A} = \widetilde{\mathbf{L}}_0\widetilde{\mathbf{U}}_0 = \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ 0 & -2 \end{pmatrix}, \quad \widetilde{\mathbf{A}}_1 := \widetilde{\mathbf{U}}_0\widetilde{\mathbf{L}}_0 = \begin{pmatrix} 5 & 1 \\ -6 & -2 \end{pmatrix};$$

$$\widetilde{\mathbf{A}}_1 = \widetilde{\mathbf{L}}_1\widetilde{\mathbf{U}}_1 = \begin{pmatrix} 1 & 0 \\ -1.2 & 1 \end{pmatrix}\begin{pmatrix} 5 & 1 \\ 0 & -0.8 \end{pmatrix}, \quad \widetilde{\mathbf{A}}_2 := \widetilde{\mathbf{U}}_1\widetilde{\mathbf{L}}_1 = \begin{pmatrix} 3.8 & 1 \\ 0.96 & -0.8 \end{pmatrix};$$

$$\widetilde{\mathbf{A}}_2 = \widetilde{\mathbf{L}}_2\widetilde{\mathbf{U}}_2 = \begin{pmatrix} 1 & 0 \\ 0.2526... & 1 \end{pmatrix}\begin{pmatrix} 3.8 & 1 \\ 0 & -1.0526... \end{pmatrix},$$

$$\widetilde{\mathbf{A}}_3 := \widetilde{\mathbf{U}}_2\widetilde{\mathbf{L}}_2 = \begin{pmatrix} 4.0526... & 1 \\ -0.2659... & -1.0526... \end{pmatrix}.$$

Диагонали матриц $\mathbf{A}_k$ и $\widetilde{\mathbf{A}}_k$, несущие приближения к собственным числам $\mathbf{A}$, при одних и тех же значениях k полностью совпадают, но

---

*) Для LU-разложения здесь поочередно используются формулы

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj} \quad \text{при } i \le j, \qquad l_{ij} = \frac{1}{u_{jj}}\left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj}\right) \quad \text{при } i > j$$

во втором случае не так заметно стремление к нулю поддиагональных элементов, хотя, относительная скорость убывания модулей наддиагональных элементов $\mathbf{A}_k$ и поддиагональных элементов $\widetilde{\mathbf{A}}_k$ примерно одинакова.

Одним из серьезных факторов, ограничивающих сферу применения LU-алгоритмов, является их недостаточно хорошая численная устойчивость (улучшение этого параметра путем перестановок строк и столбцов сильно отражается на экономичности метода). Этот фактор может играть особенно существенную роль на фоне возможной неустойчивости самой несимметричной проблемы собственных значений.

Ярким примером матрицы, для которой задача нахождения собственных чисел является неустойчивой, служит $n \times n$-матрица [138]

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 0 & 0 \end{pmatrix},$$

имеющая число 0 собственным значением n-й кратности. Введем возмущение ε в левый нижний элемент матрицы $\mathbf{A}$. Характеристическим уравнением для возмущенной матрицы

$$\mathbf{A}_\varepsilon = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 0 & 1 \\ \varepsilon & 0 & 0 & \dots & 0 & 0 \end{pmatrix}$$

служит уравнение

$$\begin{vmatrix} -\lambda & 1 & 0 & \dots & 0 & 0 \\ 0 & -\lambda & 1 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & -\lambda & 1 \\ \varepsilon & 0 & 0 & \dots & 0 & -\lambda \end{vmatrix} = 0.$$

Раскрывая определитель по элементам первого столбца, получаем:

$$-\lambda\cdot\begin{vmatrix}-\lambda & 1 & \dots & 0 & 0\\ \cdot & \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & \dots & -\lambda & 1\\ 0 & 0 & \dots & 0 & -\lambda\end{vmatrix}+(-1)^{n+1}\varepsilon\cdot\begin{vmatrix}1 & 0 & \dots & 0 & 0\\ -\lambda & 1 & \dots & 0 & 0\\ \cdot & \cdot & \cdot & \cdot & \cdot\\ 0 & 0 & \dots & -\lambda & 1\end{vmatrix}=0 \iff$$

$$\iff \quad -\lambda(-\lambda)^{n-1}+(-1)^{n+1}\varepsilon=0 \quad \iff \quad \lambda^n-\varepsilon=0.$$

Следовательно, матрица $\mathbf{A}_\varepsilon$ имеет n различных, в общем, комплексных собственных значений $\lambda_i=\sqrt[n]{\varepsilon}$, $i=1,2,\dots,n$. Если взять, например, $n=100$, а $\varepsilon=10^{-100}$, то $|\lambda_i|=0.1$, т.е. чрезвычайно малое, неощутимое для вычислительной машины искажение всего одного элемента данной специфической матрицы приводит к существенному изменению ее спектра.

Разумеется, большинство важных в приложениях задач на собственные значения не так плохи. Однако, обозначив и, возможно, намеренно утрировав проблему, призовем читателя к осторожности в применениях уже рассмотренных методов и интерпретации их результатов, а также к пониманию необходимости построения более устойчивых методов численного решения несимметричных спектральных алгебраических задач (см. следующий параграф). Численная устойчивость всех описываемых здесь методов подробно изучается в [179].

4.6. QR-АЛГОРИТМ

В идейном плане от схематично описанного выше LU-алгоритма мало чем отличается так называемый ***QR-алгоритм****) [3, 13, 42, 43, 75, 138, 141, 179]. При $k=0,1,2,\dots$, начиная с $\mathbf{A}_0:=\mathbf{A}$, здесь строят последовательность матриц $(\mathbf{A}_k)$ по формулам

$$\mathbf{A}_k=\mathbf{Q}_k\mathbf{R}_k,\quad \mathbf{A}_{k+1}=\mathbf{R}_k\mathbf{Q}_k, \tag{4.36}$$

*) Этот метод предложен почти одновременно российским математиком Кублановской (1961 г.) и англичанином Фрэнсисом (1962 г.).

первая из которых означает разложение матрицы $\mathbf{A}_k$ в произведение ортогональной $\mathbf{Q}_k$ и правой треугольной $\mathbf{R}_k$ (такое разложение существует для любой квадратной матрицы [13]), а вторая — перемножение полученных в результате факторизации $\mathbf{A}_k$ матриц $\mathbf{Q}_k$ и $\mathbf{R}_k$ в обратном порядке.

Аналогично предыдущему (см. § 4.5) на основе свойства ортогональных матриц $\mathbf{Q}_k^T=\mathbf{Q}_k^{-1}$ в соответствии с (4.36) можно записать представление данной матрицы $\mathbf{A}$ в виде

$$\mathbf{A}=\mathbf{Q}_0\mathbf{Q}_1\dots\mathbf{Q}_{k-1}\mathbf{Q}_k\mathbf{A}_{k+1}\mathbf{Q}_k^T\mathbf{Q}_{k-1}^T\dots\mathbf{Q}_1^T\mathbf{Q}_0^T$$

или, иначе,

$$\mathbf{A}=(\mathbf{Q}_0\mathbf{Q}_1\dots\mathbf{Q}_{k-1}\mathbf{Q}_k)\mathbf{A}_{k+1}(\mathbf{Q}_0\mathbf{Q}_1\dots\mathbf{Q}_k)^{-1}. \tag{4.37}$$

Следовательно, любая из матриц последовательности $(\mathbf{A}_k)$ ортогонально подобна матрице $\mathbf{A}$.

При определенных ограничениях, одним из которых опять выступает требование, чтобы матрица $\mathbf{A}$ не имела равных по модулю собственных значений, генерируемая процессом (4.36) последовательность матриц $(\mathbf{A}_k)$ сходится к матрице правой треугольной формы с диагональю из собственных чисел. Скорость обнуления поддиагональных частей матриц $\mathbf{A}_k$ линейна и зависит, как и во многих ранее рассмотренных методах, от отношений $|\lambda_i|/|\lambda_j|$ при $i>j$ (по-прежнему считаем $|\lambda_1|>|\lambda_2|>\dots>|\lambda_n|$). Наличие комплексно сопряженных пар собственных чисел у данной вещественной матрицы $\mathbf{A}$ не является, вообще говоря, препятствием для применения QR-алгоритма; просто в этом случае предельной матрицей для последовательности $(\mathbf{A}_k)$ будет матрица квазитреугольного (иначе, блочно-треугольного) вида. Каждой комплексной паре собственных чисел в такой матрице будет соответствовать диагональный 2×2-блок, причем сходимость здесь наблюдается по форме матрицы, а не поэлементно (т.е. элементы внутри этих блоков могут изменяться без видимой зависимости от k при сохранении неизменными их собственных чисел).

Обычно QR-алгоритм (4.36) применяют не к исходной матрице $\mathbf{A}$, а к подобной ей ***правой почти треугольной матрице*** $\mathbf{B}$,

называемой также ***матрицей Хессенберга****), вида

$$\mathbf{B}=\begin{pmatrix} b_{11} & b_{12} & \dots & b_{1,n-1} & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2,n-1} & b_{2n} \\ 0 & b_{32} & \dots & b_{3,n-1} & b_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & b_{n-1,n-1} & b_{n-1,n} \\ 0 & 0 & \dots & b_{n,n-1} & b_{nn} \end{pmatrix}.$$

В основе преобразования **A** к виду **B** (определяемому условием $b_{ij}=0$ при $j<i-1$) лежит ***преобразование Хаусхолдера*** или, иначе, ***преобразование отражения***, осуществляемое с помощью ***матрицы отражения (Хаусхолдера)***

$$\mathbf{H}=\mathbf{E}-2\mathbf{W}\mathbf{W}^T,$$

где **W** — произвольный вектор-столбец, но такой, что его евклидова норма равна единице. В силу этого требования к вектору **W**, выполняется равенство

$$\|\mathbf{W}\|_2^2=(\mathbf{W},\mathbf{W})=\mathbf{W}^T\mathbf{W}=1,$$

и с учетом симметричности матрицы **H**, вытекающей из симметричности матрицы

$$\mathbf{W}\mathbf{W}^T=\begin{pmatrix} W_1 \\ W_2 \\ \dots \\ W_n \end{pmatrix}\cdot(W_1;\ W_2;\ \dots;\ W_n)=\begin{pmatrix} W_1^2 & W_1W_2 & \dots & W_1W_n \\ W_2W_1 & W_2^2 & \dots & W_2W_n \\ \dots & \dots & \dots & \dots \\ W_nW_1 & W_nW_2 & \dots & W_n^2 \end{pmatrix},$$

имеем:

$$\mathbf{H}\mathbf{H}^T=\mathbf{H}^2=\mathbf{E}-4\mathbf{W}\mathbf{W}^T+4\mathbf{W}\mathbf{W}^T\mathbf{W}\mathbf{W}^T=\mathbf{E}.$$

Следовательно, матрица отражения ортогональна, и ,значит, матрицы **A** и **B**, связанные соотношением

$$\mathbf{B}=\mathbf{HAH}\quad\left(=\mathbf{HAH}^T=\mathbf{HAH}^{-1}\right),$$

являются подобными*).

Теперь нетрудно сообразить, как нужно распорядиться свободой задания элементов векторов **W** при построении матриц отражения, чтобы за конечное число шагов преобразований Хаусхолдера произвольно заданную матрицу **A** привести к форме Хессенберга **B**.

А именно, можно показать, что начатым с $\mathbf{B}_1:=\mathbf{A}$ процессом

$$\mathbf{B}_{m+1}=\mathbf{H}_m\mathbf{B}_m\mathbf{H}_m,\quad m=1,2,\dots,n-2, \tag{4.38}$$

где $\mathbf{H}_m=\mathbf{E}-2\mathbf{W}_m\mathbf{W}_m^T$, данная $n\times n$-матрица **A** за $n-2$ шага будет приведена к виду **B**, т.е. матрица $\mathbf{B}:=\mathbf{B}_{n-1}$ подобна **A**, если задающие матрицы Хаусхолдера $\mathbf{H}_m$ векторы $\mathbf{W}_m$ по данной матрице **A** строить следующим образом**).

При $m=1$ вектор $\mathbf{W}_1$ определяется равенством

$$\mathbf{W}_1^T=\mu_1(0;a_{21}-s_1;a_{31};\dots;a_{n1}), \tag{4.39}$$

где $s_1=\operatorname{sign}(-a_{21})\cdot\sqrt{\sum_{i=2}^{n}a_{i1}^2}$, $\mu_1=\dfrac{1}{\sqrt{2s_1(s_1-a_{21})}}$. Такое задание $\mathbf{W}_1$ обеспечивает ортогональность симметричной матрицы

$$\mathbf{H}_1=\mathbf{E}-2\mathbf{W}_1\mathbf{W}_1^T$$

и одновременное получение с ее помощью нужных $n-2$ нулей в первом столбце матрицы

$$\mathbf{B}_2=\mathbf{H}_1\mathbf{B}_1\mathbf{H}_1\quad(=\mathbf{H}_1\mathbf{A}\mathbf{H}_1).$$

Вектор $\mathbf{W}_2$ по матрице $\mathbf{B}_2$ строится совершенно аналогично, только фиксируются нулевыми не одна, а две первые его координаты, и определяющую роль играют теперь не первый, а второй столбец матрицы $\mathbf{B}_2$ и его третий элемент. При этом у матрицы $\mathbf{B}_3=\mathbf{H}_2\mathbf{B}_2\mathbf{H}_2$ окажется $n-3$ нулевых элемента во втором столбце и сохранятся полученные на предыдущем шаге нули в первом столбце.

---

*) Герхард Хессенберг (1874–1925) — немецкий математик. Иногда почти треугольной формой матрицы называют упомянутые матрицы блочно-треугольного вида [138].

*) Матрица отражения обладает рядом других интересных свойств; в частности, ее название связано с тем, что линейное преобразование, осуществляемое такой матрицей, оставляет без изменений векторы, ортогональные вектору **W**, а коллинеарные ему векторы переводит в противоположные («отражает»).

**) Логику таких построений, основанную на сохранении евклидовой нормы столбца, см., например, в [179].

Этот процесс очевидным образом может быть продолжен до исчерпания и без особого труда может быть описан общими формулами типа формул (4.39), для чего нужно лишь ввести обозначения для элементов последовательности матриц $\mathbf{B}_m$ (например, с помощью верхних индексов m).

Рассмотрим на простом числовом примере, как приводится матрица к форме Хессенберга, когда для этого требуется только один шаг преобразований Хаусхолдера.

Пример 4.5. Дана матрица $\mathbf{A}=\begin{pmatrix} 5 & 1 & -3 \\ 3 & 0 & -2 \\ -4 & -1 & 1 \end{pmatrix}$. Найти матрицу $\mathbf{B}$, подобную матрице $\mathbf{A}$ и имеющую форму Хессенберга.

Решение проводим по формулам (4.38), (4.39) при $n=3$, которые для данного случая можно записать так (в естественном для выполнения порядке):

$$s=\operatorname{sign}(-a_{21})\cdot\sqrt{a_{21}^2+a_{31}^2}\,; \quad \mu=\frac{1}{\sqrt{2s(s-a_{21})}}\,;$$

$$\mathbf{W}^T=\mu(0;\ a_{21}-s;\ a_{31}); \quad \mathbf{H}=\mathbf{E}-2\mathbf{W}\mathbf{W}^T; \quad \mathbf{B}=\mathbf{HAH}.$$

Имеем:

$$s=-\sqrt{3^2+(-4)^2}=-5; \quad \mu=\frac{1}{\sqrt{2(-5)(-5-3)}}=\frac{1}{4\sqrt{5}};$$

$$2\mathbf{W}\mathbf{W}^T=\frac{1}{40}\begin{pmatrix} 0 \\ 8 \\ -4 \end{pmatrix}(0;\,8;\,-4)=\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1.6 & -0.8 \\ 0 & -0.8 & 0.4 \end{pmatrix};$$

$$\mathbf{H}=\begin{pmatrix} 1 & 0 & 0 \\ 0 & -0.6 & 0.8 \\ 0 & 0.8 & 0.6 \end{pmatrix}; \quad \mathbf{HA}=\begin{pmatrix} 5 & 1 & -3 \\ -5 & -0.8 & 2 \\ 0 & -0.6 & -1 \end{pmatrix};$$

$$\mathbf{B}=\begin{pmatrix} 5 & -3 & -1 \\ -5 & 2.08 & 0.56 \\ 0 & -0.44 & -1.08 \end{pmatrix}.$$

Замечание 4.9. При решении с и м м е т р и ч н ы х проблем собственных значений методом вращений на первой стадии решения также часто применяют преобразования Хаусхолдера. Абсолютно те же формулы (4.38), (4.39) приведут симметричную матрицу $\mathbf{A}$ к подобной ей матрице $\mathbf{B}$ т р е х д и а г о н а л ь н о г о вида (частному случаю формы Хессенберга), что значительно повышает эффективность последующих преобразований плоских вращений Якоби.

Замечание 4.10. Если начать процесс построения матриц отражения $\mathbf{H}_m$ не по формулам (4.39), а по аналогичным им, но с ключевым элементом a_{11}, т.е. полагая

$$\mathbf{W}_1^T=\mu_1(a_{11}-s;\ a_{21};\ \ldots;\ a_{n1}),$$

$$s_1=\operatorname{sign}(-a_{11})\cdot\sqrt{\sum_{i=1}^{n}a_{i1}^2}\,, \quad \mu_1=\frac{1}{\sqrt{2s_1(s_1-a_{11})}}\,,$$

то за $n-1$ шаг ортогональных преобразований

$$\mathbf{R}_{m+1}=\mathbf{H}_m\mathbf{R}_m; \quad m=1,2,\ldots,n-1; \quad \mathbf{R}_1:=\mathbf{A}$$

можно получить разложение матрицы $\mathbf{A}$ в произведение ортогональной $\mathbf{H}:=\mathbf{H}_{n-1}\ldots\mathbf{H}_2\mathbf{H}_1$ и правой треугольной $\mathbf{R}:=\mathbf{R}_n$, так как результирующее равенство $\mathbf{R}=\mathbf{HA}$ равносильно равенству $\mathbf{A}=\mathbf{HR}$ в силу ортогональности и симметричности $\mathbf{H}$. Приведение матрицы $\mathbf{A}$ такими преобразованиями к треугольному виду составляет основу ***метода отражений решения линейных алгебраических систем***. Из равносильности равенств $\mathbf{Ax}=\mathbf{b}$, $\mathbf{HAx}=\mathbf{Hb}$ и $\mathbf{Rx}=\mathbf{Hb}$ легко понять, что для решения СЛАУ этим методом нужно над вектором свободных членов $\mathbf{b}$ выполнять те же преобразования, что и над матрицей коэффициентов $\mathbf{A}$, после чего нужно будет сделать только обратный ход, как в методе Гаусса. Метод отражений решения СЛАУ в полтора раза экономичнее метода вращений (§ 2.7) и практически не уступает последнему по устойчивости к накоплению ошибок округлений. Более подробно об этом методе см., например, в [54].

Вообще говоря, весь QR-алгоритм (4.36) от начала и до конца может быть построен на базе описанной выше процедуры преобразований Хаусхолдера, направленной сразу на триангуляризацию в соответствии с замечанием 4.10. Однако такой подход значительно сужает границы применимости алгоритма и ухудшает его скоростные качества.

Обычно на втором этапе применяют другое ортогональное преобразование — ***преобразование плоских вращений Гивенса***. Определяющая эти преобразования матрица $\mathbf{G}_{ij}=(g_{ml})_{m,l=1}^{n}$ при фиксированных i,j — индексах ключевого элемента преобразуемой матрицы — имеет точно такую же структуру, как и матрица плоских вращений Якоби $\mathbf{T}_{ij}$ (ср.(4.27)), только здесь, следуя [138], двумерную подматрицу из элементов, стоящих на пересечении i-х и j-х строк и столбцов, возьмем в виде

$$\hat{\mathbf{G}}_{ij}=\begin{pmatrix} s & c \\ -c & s \end{pmatrix}.$$

Как и прежде, числа s и c связываем соотношением $s^2+c^2=1$ (это позволяет интерпретировать их как синус и косинус некоторо-

го угла θ), обеспечивающим ортонормированность матриц $\mathbf{G}_{ij}$.

Первый полный шаг преобразования Гивенса, применяемого к матрице Хессенберга $\mathbf{B}$ n-го порядка в рамках QR-алгоритма (4.36), состоит из $n-1$ элементарных подшагов, имеющих целью последовательное обнуление поддиагональных элементов в столбцах от первого до $(n-1)$-го. В результате этого получается разложение матрицы $\mathbf{B}$ в произведение ортогональной и треугольной, что требуется первой формулой (4.36) при $k=1$, $\mathbf{A}_1 := \mathbf{B}$.

Чтобы определить s и c на[*] первом промежуточном шаге, рассмотрим произведение матриц

$$\mathbf{G}_1\mathbf{B} = \begin{pmatrix} s & c & 0 & \dots & 0 \\ -c & s & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & b_{13} & \dots & b_{1n} \\ b_{21} & b_{22} & b_{23} & \dots & b_{2n} \\ 0 & b_{32} & b_{33} & \dots & b_{3n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & b_{nn} \end{pmatrix} =$$

$$= \begin{pmatrix} sb_{11}+cb_{21} & sb_{12}+cb_{22} & sb_{13}+cb_{23} & \dots & sb_{1n}+cb_{2n} \\ -cb_{11}+sb_{21} & -cb_{12}+sb_{22} & -cb_{13}+sb_{23} & \dots & -cb_{1n}+sb_{2n} \\ 0 & b_{32} & b_{33} & \dots & b_{3n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & b_{nn} \end{pmatrix}.$$

Беря $c=\cos\theta$ и $s=\sin\theta$ такими, что $\operatorname{tg}\theta = \dfrac{b_{11}}{b_{21}}$, будем иметь $b_{11}\cos\theta = b_{21}\sin\theta$, т.е. $-c_1 b_{11} + s b_{21} = 0$. Значит, результат первого промежуточного шага — матрица $\mathbf{B}_1 = \mathbf{G}_1\mathbf{B}$, получающаяся при таких c и s, не будет содержать ненулевых элементов под диагональю в первом столбце.

Второй промежуточный шаг совершается аналогично: матрица $\mathbf{B}_2 = \mathbf{G}_2\mathbf{B}_1$ получается из предыдущей $\mathbf{B}_1$ с помощью матрицы Гивенса $\mathbf{G}_2$, отличающейся от $\mathbf{G}_1$ тем, что подматрица $\begin{pmatrix} s & c \\ -c & s \end{pmatrix}$ смещается на одну позицию вдоль диагонали и угол поворота подбирается так, чтобы в матрице $\mathbf{B}_2$ обнулить элемент $b_{32}^{(2)}$.

[*] Так как в каждом столбце матрицы Хессенберга нужно «убивать» только по одному элементу, то в данном случае матрицы элементарных вращений Гивенса можно помечать только одним индексом.

Продолжая этот процесс преобразований Гивенса далее, в итоге получим правую треугольную матрицу

$$\mathbf{B}_{n-1} = \mathbf{G}_{n-1}\mathbf{G}_{n-2}\dots\mathbf{G}_2\mathbf{G}_1\mathbf{B}.$$

Последнее равенство можно переписать в виде

$$(\mathbf{G}_{n-1}\mathbf{G}_{n-2}\dots\mathbf{G}_2\mathbf{G}_1)^{-1}\mathbf{B}_{n-1} = \mathbf{B},$$

который позволяет считать выполненным требуемое в (4.36) при $k=1$ разложение

$$\mathbf{B} = \mathbf{Q}_1\mathbf{R}_1,$$

где $\mathbf{Q}_1 := (\mathbf{G}_{n-1}\mathbf{G}_{n-2}\dots\mathbf{G}_2\mathbf{G}_1)^{-1} = \mathbf{G}_1^T\dots\mathbf{G}_{n-1}^T$ — ортогональная, а $\mathbf{R}_1 := \mathbf{B}_{n-1}$ — правая треугольная матрицы. При этом матрица

$$\mathbf{A}_2 := \mathbf{R}_1\mathbf{Q}_1 = (\mathbf{G}_{n-1}\dots\mathbf{G}_1)\mathbf{B}(\mathbf{G}_{n-1}\dots\mathbf{G}_1)^{-1}, \qquad (4.40)$$

являющаяся результатом первого полного шага QR-алгоритма (примененного к $\mathbf{B}$), сохраняет не только спектр данной матрицы, но и форму Хессенберга [43, 138], благодаря чему приведение исходной матрицы $\mathbf{A}$ к почти треугольному виду $\mathbf{B}$ достаточно сделать только один раз.

Очевидно, скалярные параметры $c_j = \cos\theta_j$ и $s_j = \sin\theta_j$ матриц Гивенса $\mathbf{G}_j$, с помощью которых осуществляется переход от матрицы Хессенберга $\mathbf{B} = (b_{ij})_{i,j=1}^n$ «транзитом» через матрицы Хессенберга $\mathbf{B}_j = \left(b_{im}^{(j)}\right)_{i,\, m=1}^n$ к матрице Хессенберга $\mathbf{A}_2$, можно вычислять на j-м промежуточном шаге ($j = 1, 2, \dots, n-1$) по формулам

$$c_j = \frac{1}{\sqrt{1+t_j^2}}, \quad s_j = t_j c_j,$$

где

$$t_j = \frac{b_{jj}^{(j-1)}}{b_{j+1,j}^{(j-1)}} \; (= \operatorname{tg}\theta_j), \quad b_{ij}^{(0)} := b_{ij}$$

(если знаменатель в выражении t_j равен нулю или по модулю меньше некоторого существенно малого порогового значения, то можно считать $c_j = 0$, $s_j = 1$, т.е. $\mathbf{G}_j := \mathbf{E}$).

Пример 4.6. Преобразованиями Гивенса выполним один шаг QR-алгоритма для матрицы $\mathbf{B} = \begin{pmatrix} 5 & -3 & -1 \\ -5 & 2.08 & 0.56 \\ 0 & -0.44 & -1.08 \end{pmatrix}$, полученной в результате преобразований Хаусхолдера в предыдущем примере.

При $j=1$ последовательно находим:

$$t_1 = \frac{5}{-5} = -1, \quad c_1 = \frac{1}{\sqrt{1+(-1)^2}} = \frac{1}{\sqrt{2}}, \quad s_1 = -\frac{1}{\sqrt{2}};$$

$$\mathbf{G}_1 = \begin{pmatrix} -0.5\sqrt{2} & 0.5\sqrt{2} & 0 \\ -0.5\sqrt{2} & -0.5\sqrt{2} & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\mathbf{B}_1 = \mathbf{G}_1\mathbf{B} = \begin{pmatrix} -5\sqrt{2} & 2.54\sqrt{2} & 0.78\sqrt{2} \\ 0 & 0.46\sqrt{2} & 0.22\sqrt{2} \\ 0 & -0.44 & -1.08 \end{pmatrix}.$$

При $j=2$ вычисляем (округляя до 10^{-6}):

$$t_2 = \frac{0.46\sqrt{2}}{-0.44} = -1.478496, \quad c_2 = 0.560248, \quad s_2 = -0.828325;$$

значит,

$$\mathbf{G}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -0.828325 & 0.560248 \\ 0 & -0.560248 & -0.828325 \end{pmatrix},$$

$$\mathbf{R}_1 := \mathbf{B}_2 = \mathbf{G}_2\mathbf{B}_1 = \begin{pmatrix} -7.071068 & 3.592102 & 1.103087 \\ 0 & -0.785366 & -0.862782 \\ 0 & 0 & 0.720283 \end{pmatrix}.$$

Следовательно,

$$\mathbf{Q}_1 = \mathbf{G}_1^T\mathbf{G}_2^T = \begin{pmatrix} -0.707107 & -0.585714 & -0.396155 \\ 0.585714 & -0.396155 & 0.396155 \\ 0 & 0.560248 & -0.828325 \end{pmatrix},$$

и, согласно (4.40), матрица

$$\mathbf{A}_2 = \mathbf{R}_1\mathbf{Q}_1 = \begin{pmatrix} 7.103946 & 3.336597 & 3.310554 \\ -0.460000 & -0.172245 & 0.403537 \\ 0 & 0.403537 & -0.596628 \end{pmatrix}$$

есть искомый результат первого полного шага QR-алгоритма. Она имеет те же собственные числа, что $\mathbf{B}$ и $\mathbf{A}$, сохраняет форму Хессенберга и модули ее поддиагональных элементов меньше, чем у матрицы $\mathbf{B}$, т.е. она более близка к подобной $\mathbf{B}$ матрице треугольного вида, на диагонали которой должны быть собственные числа данной матрицы ($\lambda_1 \approx 7.693$, $\lambda_2 \approx -1.205$, $\lambda_3 \approx -0.435$).

Замечание 4.11. Весь QR-алгоритм можно было бы построить на базе одних только преобразований Гивенса, т.е. не приводя исходную матрицу $\mathbf{A}$ к форме Хессенберга (или к трехдиагональному виду, если $\mathbf{A}$ симметрична) другими преобразованиями. В таком случае стала бы заметной разница между преобразованиями Якоби и Гивенса. Суть этой разницы в следующем: если для преобразований Якоби понятия «ключевой элемент» и «обреченный элемент» совпадают, то для преобразований Гивенса это, вообще говоря, не так. В общем случае при вращениях Гивенса угол поворота θ в фиксированной индексами i, j плоскости вращения подбирается так, чтобы аннулировать какой-нибудь элемент, стоящий либо в одном столбце, либо в одной строке с ключевым элементом a_{ij}. Такие преобразования теряют свойство минимальности суммы квадратов внедиагональных элементов, имевшее место в преобразованиях Якоби для симметричных матриц, но позволяют (Гивенс, 1954г. [179]) привести симметричную матрицу к трехдиагональному виду существенно быстрее, чем это требуется для выполнения одного цикла преобразований в методе вращений Якоби*). Приведение несимметричных матриц к форме Хессенберга методом Гивенса требует большего числа арифметических операций, чем это нужно для такого приведения методом Хаусхолдера, поэтому обычно для этих целей отдают предпочтение последнему.

Приведенных выше сведений, в принципе, вполне достаточно, чтобы находить QR-алгоритмом все хорошо отделимые вещественные собственные числа вещественных матриц, реализуя равенство (4.37) при некотором k, хотя здесь и нет должного для этого обоснования**). Однако в такой непосредственной форме QR-алгоритм не применяется ввиду его медленной сходимости. Для ускорения сходимости в процесс (4.36) вводят ***сдвиги***,

---

*) Для вычисления собственных значений трехдиагональных матриц разработан не рассматриваемый здесь довольно эффективный ***метод бисекций*** [29, 42, 179], который также можно считать одним из способов локализации собственных чисел.

**) Доказательство сходимости QR-алгоритма и его модификаций см. в [75, 179]. В [141] изучается сходимость принятого там за основу QL-алгоритма.

о роли которых немало говорилось при изучении метода обратных итераций (см. § 4.3). При выборе параметров сдвигов в QR-алгоритме учитывается подмеченная ранее целесообразность использования для этого приближений к собственным числам. В данном случае принимается во внимание, что формируемая QR-алгоритмом последовательность матриц $\mathbf{A}_k$ в пределе дает правую треугольную матрицу с диагональю из собственных чисел (когда все собственные числа матрицы $\mathbf{A}$ вещественны). Следовательно, есть основания утверждать, что последовательность элементов $a_{nn}^{(k)}$ может рассматриваться при $k=1, 2, \ldots$ как последовательность приближений к какому-то определенному собственному числу матрицы $\mathbf{A}$ и служить соответствующей последовательностью параметров переменных сдвигов, ускоряющих процесс обнуления поддиагональных элементов.

Таким образом, по крайней мере, в случае, когда данная матрица $\mathbf{A}$ имеет только вещественные собственные значения, QR-алгоритм будет сходиться более быстро (квадратично), если при каждом k преобразования Гивенса применять не к матрице $\mathbf{A}_k$, а к матрице $\widetilde{\mathbf{A}}_k := \mathbf{A}_k - a_{nn}^{(k)}\mathbf{E}$. При этом каждый раз спектр матрицы смещается на величину произведенного сдвига (см. свойство 4.2 в § 4.1), что может учитываться двояко: либо параметры сдвигов суммируются и затем сумма прибавляется к найденным в итоге значениям, либо каждый раз делается обратный сдвиг, т.е. вместо формул (4.36) используются формулы:

$$\mathbf{A}_k - a_{nn}^{(k)}\mathbf{E} = \mathbf{Q}_k\mathbf{R}_k$$

(QR-факторизация матрицы $\mathbf{A}_k - a_{nn}^{(k)}E$),

$$\mathbf{A}_{k+1} = \mathbf{R}_k\mathbf{Q}_k + a_{nn}^{(k)}\mathbf{E}$$

(перемножение в обратном порядке и обратный сдвиг).

Важно учесть, что большая экономия вычислительных затрат при реализации QR-алгоритма получается в результате включения сюда ***процедуры исчерпывания***. Нетрудно показать, что если на каком-то шаге $k=k_0$ последняя строка матрицы n-го порядка $\mathbf{A}_{k_0}$ стала нулевой (с некоторой точностью), то это позволяет считать $a_{nn}^{(k_0)}$ собственным числом, а другие ее собственные значения находить, работая далее только с подматрицей $(n-1)$-го порядка, получающейся из $\mathbf{A}_{k_0}$ отбрасыванием последних строки и столбца.

Если у данной матрицы возможны комплексные собственные значения, применяют более сложный процесс двойных сдвигов [75, 138, 179], позволяющий преодолеть ситуацию, когда в конце диагоналей матриц $\mathbf{A}_k$ «прорисовывается» 2×2-блок, соответствующий паре комплексно сопряженных собственных чисел.

Нахождение собственных векторов в рамках QR-алгоритма (и других методов, основанных на асимптотической триангуляризации) не является такой простой задачей, как это было в методе вращений Якоби для симметричных матриц, диагонализируемых в процессе ортогональных преобразований. Однако при известном собственном числе соответствующий ему собственный вектор эффективно может быть найден рассмотренным в § 4.3 методом обратных итераций (см. формулы (4.19),(4.20)). При этом обратные итерации обычно применяются не к исходной матрице $\mathbf{A}$, а к матрице $\mathbf{B}$, подобной $\mathbf{A}$ и имеющей форму Хессенберга. Если приведение $\mathbf{A}$ к виду $\mathbf{B}$ выполнялось преобразованиями Хаусхолдера (4.38), то $\mathbf{B}=\mathbf{HAH}$, где $\mathbf{H}$ — результирующая матрица $n-2$ элементарных вращений, и значит, согласно свойству 4.7 (§ 4.4), найдя собственный вектор $\mathbf{y}$ матрицы $\mathbf{B}$, искомый собственный вектор $\mathbf{x}$ данной матрицы $\mathbf{A}$ получаем равенством $\mathbf{x}=\mathbf{Hy}$.

УПРАЖНЕНИЯ

4.1. Дана матрица $\mathbf{A}=\begin{pmatrix} 30 & -12 & 53 \\ -42 & 19 & -78 \\ -28 & 12 & -51 \end{pmatrix}$.

А) Степенным методом найдите несколько последовательных приближений к доминирующему собственному числу матрицы $\mathbf{A}$ и к соответствующему собственному вектору. Зная, что искомое собственное число есть $\lambda_1=-5$, проверьте, насколько эффективно здесь применение Δ^2-процесса Эйткена (см. замечание 4.6).

Б) Методом обратных итераций найдите младшую собственную пару $\{\lambda_3, \mathbf{x}_3\}$ данной матрицы $\mathbf{A}$. Можно ли утверждать, что $\lambda_3=\Lambda+\lambda_1$, где Λ — наибольшее по модулю собственное число матрицы $\mathbf{A}-\lambda_1\mathbf{E}$?

4.2. В условиях примера 4.1 начните SP-алгоритм (§ 4.2) с вектора $\mathbf{y}^{(0)}=(1;\ 1)^T$. Что получено после выполнения одного полного цикла алгоритма? Дайте объяснение результату.

4.3. Найдя грубые приближения к собственным числам матрицы

$$\mathbf{A} = \begin{pmatrix} 5 & 2 & -3 \\ 4 & 5 & -4 \\ 6 & 4 & -4 \end{pmatrix}$$

степенным методом, уточните эти значения обратными итерациями со сдвигами (см. (4.24)).

4.4. А) Проанализируйте сходимость степенного метода в случае, когда λ_1 — кратное вещественное наибольшее по модулю собственное число n-мерной матрицы простой структуры (см. формулы (4.10), (4.11)). Как можно найти все соответствующие ему собственные векторы в зависимости от показателя кратности?

Б) Что можно сказать о поведении последовательности отношений (4.10), если $\lambda_1 = -\lambda_2$, и $|\lambda_2| > |\lambda_3| \geq \ldots \geq |\lambda_n|$ $(\lambda_i \in R)$?

В) Рассмотрите и объясните поведение степенного метода в случае, когда данная матрица $\mathbf{A}$ — диагональная.

4.5. Найдите все собственные пары матрицы

$$\mathbf{A} = \begin{pmatrix} 4 & 2 & -1 \\ 2 & 4 & 1 \\ -1 & 1 & 3 \end{pmatrix}:$$

а) методом скалярных произведений (для нахождения второй собственной пары используйте формулы (4.14), (4.15));

б) RQI-алгоритмом, начиная его с различных векторов.

4.6. Методом вращений Якоби найдите собственные пары матрицы $\mathbf{A}$, если:

а) $\mathbf{A} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$; б) $\mathbf{A} = \begin{pmatrix} 6 & -2 & 2 \\ -2 & 5 & 0 \\ 2 & 0 & 7 \end{pmatrix}$.

4.7. Сравните два подхода к нахождению всех собственных чисел матрицы

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

методом вращений Якоби:

А) применяя его непосредственно к данной матрице;

Б) предварительно приведя ее к трехдиагональному виду преобразованиями Хаусхолдера.

4.8. Для нахождения собственных пар симметричных положительно определенных матриц постройте LU-алгоритм на базе $\mathrm{U}^T\mathrm{U}$ (или LL^T)-разложения Холецкого (см. гл. 2). Опробуйте его на матрицах

$$\mathbf{A} = \begin{pmatrix} 5 & -2 \\ -2 & 2 \end{pmatrix} \quad \text{и} \quad \mathbf{B} = \begin{pmatrix} 6 & 1 & 0 \\ 1 & 5 & -2 \\ 0 & -2 & 1 \end{pmatrix}.$$

Сохраняют ли получаемые на каждом шаге такого алгоритма подобные $\mathbf{B}$ матрицы трехдиагональную структуру?

4.9. Дана матрица $\mathbf{A} = \begin{pmatrix} 3 & -1 \\ -2 & 2 \end{pmatrix}$. Сделайте по три шага:

а) LU-алгоритма;

б) QR-алгоритма на основе преобразований Гивенса;

в) QR-алгоритма на основе преобразований Хаусхолдера (с учетом замечания 4.10).

Сравните полученные приближения к собственным числам матрицы $\mathbf{A}$ по точности (найдя сначала ее точные значения с помощью характеристического уравнения) и по вычислительным затратам.

4.10. Матрицу $\mathbf{A}$ из упр. 4.7 приведите к трехдиагональному виду преобразованием Гивенса (отличным от преобразования Якоби, см. замечание 4.11).

4.11. Для $n \times n$-матрицы сделайте приблизительный подсчет количества арифметических операций, приходящихся на:

а) один шаг степенного метода;

б) один шаг метода обратных итераций (без сдвигов);

в) один полный цикл метода вращений Якоби;

г) полный цикл приведения матрицы к форме Хессенберга преобразованиями Хаусхолдера.

4.12. Методом отражений решите систему

$$\begin{cases} x - 2y + z = 0, \\ 2x - y - 2z = 6, \\ 2x - y - z = 9 \end{cases}$$

(см. замечание 4.10).

4.13. Выведите формулы для преобразования $n \times n$-матрицы Хессенберга к треугольной форме плоскими поворотами на углы φ_j $(j=1, 2, \ldots, n-1)$ с помощью матриц $\mathbf{T}_j := \mathbf{T}_{j,\, j+1}$ вида (4.27). Как отличаются углы φ_j от углов θ_j, неявно присутствующих в представлении (4.40)?

ГЛАВА 5 МЕТОДЫ РЕШЕНИЯ НЕЛИНЕЙНЫХ СКАЛЯРНЫХ УРАВНЕНИЙ

Рассматриваются различные аспекты решения одномерных нелинейных уравнений. В частности, обсуждается задача локализации корней и простейшие способы сужения промежутков, их существования, даются определения основных типов сходимости итерационных последовательностей и связанных с ними порядков итерационных методов. Наряду с методами, обладающими линейной сходимостью, такими, как метод дихотомии и метод хорд, изучаются методы более высоких порядков. Наибольшее внимание здесь уделяется методу касательных Ньютона и методу секущих, рассматриваемому в качестве одной из модификаций метода Ньютона, превосходящей его по эффективности в смысле вычислительных затрат. Последний параграф главы посвящен новым методам, так называемым полюсным методам Ньютона и секущих, которые при надлежащем фиксировании параметров расширяют области применения классических методов Ньютона и секущих и в равных с ними условиях выигрывают у них в быстроте сходимости (без дополнительных вычислений значений функции).

5.1. ЛОКАЛИЗАЦИЯ КОРНЕЙ

Будем рассматривать ***задачу приближенного нахождения нулей функции одной переменной***, иначе, задачу нахождения корней уравнения вида

$$f(x) = 0, \qquad (5.1)$$

где $f:\mathbf{R}_1 \to \mathbf{R}_1$ — алгебраическая или трансцендентная функция. Такие уравнения называют ***скалярными, числовыми, конечными*** и т.п. Методам их решения посвящена обширная литература; кроме перечисленных здесь учебных и научных изданий, эту тему можно встретить почти в любом учебнике математического анализа или высшей математики, а во многих учебных пособиях по программированию даются «рецепты» решения таких

задач*). Однако не все авторы учебников по вычислительной математике считают нужным включать рассматриваемую тему самостоятельным разделом. И в этом есть резон. Действительно, многие наиболее фундаментальные методы решения скалярных уравнений можно рассматривать как частные случаи соответствующих методов решения систем нелинейных уравнений со многими неизвестными и даже, более того, нелинейных операторных уравнений (как правило, в банаховых пространствах). Такой путь изучения методов более короток, но и более труден. Учитывая, что одномерный случай более прост и легко интерпретируется геометрически (что немаловажно для нелинейных задач), а также то, что обычно теоретические результаты, полученные для скалярных уравнений, затем переносятся (не без потерь) на системы и операторные уравнения, изучаемые методы будут гораздо лучше поняты, если в них сначала как следует разобраться на объектах вида (5.1).

В общем случае, если речь идет не об отдельных достаточно узких классах уравнений, например, изучавшихся в школьном курсе математики, можно говорить лишь о п р и б л и ж е н н о м вычислении корней уравнений (5.1), т.е. таких значений аргумента $x = \xi$, при которых равенство

$$f(\xi) = 0$$

истинно. При этом под близостью приближенного значения $\bar{x}$ к корню ξ уравнения (5.1), как правило, понимают выполнение неравенства

$$|\xi - \bar{x}| < \varepsilon$$

при малых $\varepsilon > 0$, хотя часто бывает важным контролировать не абсолютную погрешность приближенного равенства $\bar{x} \approx \xi$, а относительную, т.е. величину $|\xi - \bar{x}|/|\bar{x}|$, например, когда величина $|\bar{x}|$ близка к нулю.

Нелинейная функция $f(x)$ в своей области определения $D(f) \subseteq \mathbf{R}_1$ может иметь конечное или бесконечное количество нулей или не иметь их вовсе. Большинство же методов нахожде-

*) Можно понять преподавателей, обучающих алгоритмическим языкам: методы решения скалярных уравнений предоставляют для этого благодатный материал. К сожалению, программирование формул без понимания их сути часто лишь компрометирует вычислительную математику.

ния нулей требует знания промежутков (возможно, малых), где заведомо имеется и притом единственный нуль функции*). Если такие конкретные промежутки не предоставляются постановщиком задачи, то на первый план выходят: выявление ситуации с наличием и количеством корней уравнения (5.1), нахождение области их расположения, получение отрезков, на которых имеется точно по одному корню. Иными словами, ставятся ***подзадачи существования и единственности, нахождения границ и локализации корней***. Эти подзадачи обычно решаются в комплексе средствами математического анализа. Но и численные методы здесь часто выступают в помощь математическому анализу: как будет видно из дальнейшего, многие теоремы сходимости итерационных методов можно считать локальными теоремами существования и единственности.

Для функций общего вида нет универсальных способов решения поставленных подзадач. Так что здесь, как говорится, все средства хороши.

Если функция $f(x)$ такова, что без особого труда можно построить ее график, этим следует воспользоваться, чтобы представить ситуацию с количеством и расположением нулей $f(x)$ выделяя те промежутки оси абсцисс, где график $y = f(x)$ пересекает Ox. (Знание графика много дает и для понимания поведения тех или иных процессов вычисления приближений к корням.) Может оказаться, что построение графика $y = f(x)$ вызывает затруднения, но исходное уравнение (5.1) очевидным образом представляется в виде

$$f_1(x) = f_2(x)$$

и функции $f_1(x)$ и $f_2(x)$ таковы, что легко строятся графики $y = f_1(x)$ и $y = f_2(x)$. Тогда задача определения количества корней и областей их единственности решается отслеживанием точек пересечения этих графиков и выделением на оси абсцисс тех промежутков, которым принадлежат проекции таких точек. Описанный прием называют ***графическим способом локализации*** (иначе, ***отделения***, ***изоляции***) ***корней***.

Пример 5.1. Найдем промежутки изоляции корней уравнения

$$x^2 - \sin x - 1 = 0.$$

Представив это уравнение в виде $x^2 - 1 = \sin x$, строим схематично

*) Перефразируя Конфуция, можно сказать: трудно найти корень на бесконечном промежутке, особенно, когда его там нет.

графики функций $y = x^2 - 1$ и $y = \sin x$ (рис. 5.1).

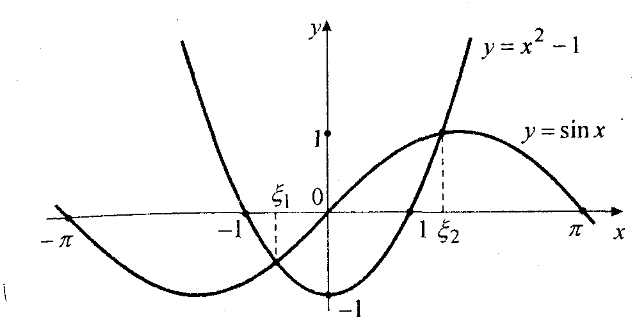


Рис. 5.1. Графическая локализация корней уравнения $x^2 - \sin x - 1 = 0$

Совместное рассмотрение графиков позволяет сделать заключение, что данное уравнение имеет два корня: $\xi_1 \in [-1, 0]$ и $\xi_2 \in [1, \pi]$.

Убедиться в том, что на данном отрезке $[a, b]$ (например, грубо определенном графическим способом) действительно имеется нуль непрерывной функции $f(x)$, можно ***аналитическим способом***, в основе которого лежит известное утверждение математического анализа.

Теорема 5.1 (***Больцано-Коши***). *Если непрерывная на отрезке* $[a, b]$ *функция* $f(x)$ *на концах его имеет противоположные знаки, т.е.*

$$f(a)f(b) < 0, \qquad (5.2)$$

то на интервале (a, b) *она хотя бы один раз обращается в нуль**).

Очевидна слабость теоремы 5.1 при ее применении к поставленной задаче: она не дает ответа на вопрос о количестве корней на отрезке $[a, b]$ в случае выполнения условия (5.2) (их может быть нечетное число) и не позволяет утверждать, что на отрезке $[a, b]$ нет корней, когда условие (5.2) не выполнено, так как в этом случае их может оказаться на $[a, b]$ четное количество (нетрудно представить себе соответствующие «картинки»

*) Допустимо считать, что a и/или b могут принимать значения $-\infty$ и/или $+\infty$ соответственно. В таком случае $f(a)$ и/или $f(b)$ следует понимать в предельном смысле, также допуская возможность бесконечных предельных значений f определенного знака.

7 Б-223

возможных поведений графиков функций на отрезке).

Результат, сформулированный в виде теоремы 5.1, можно значительно усилить, если требование непрерывности функции $f(x)$ на $[a, b]$ дополнить требованием **монотонности** ее на этом отрезке.

Теорема 5.2. *Непрерывная строго монотонная функция $f(x)$ имеет и притом единственный нуль на отрезке $[a, b]$ тогда и только тогда, когда на его концах она принимает значения разных знаков.*

Последняя теорема позволяет не только принимать, но и отвергать те или иные промежутки из области определения $D(f)$ данной функции на предмет дальнейшего поиска ее нулей, если известно о ее монотонном поведении на этих промежутках и определены знаки значений функции на их концах.

Реально установить монотонность на данном отрезке можно для **дифференцируемой** функции, потребовав знакопостоянство ее производной на всем отрезке. Для таких функций основой решения задачи локализации корней уравнения (5.1) может служить следующая теорема.

Теорема 5.3. *Пусть $f \in C^1[a, b]$. Тогда, если $f'(x)$ не меняет знак на (a, b), то условие* (5.2) *является необходимым и достаточным для того, чтобы уравнение* (5.1) *имело и притом единственный корень на $[a, b]$.*

Так как производная может менять знак только в точках, где она равна нулю или не существует, а также в граничных точках области определения функции, то в случаях (к сожалению, редких), когда уравнение $f'(x) = 0$ легко решается, вопрос о количестве и расположении корней уравнения (5.1) не вызывает трудностей.

Пример 5.2. Выясним, сколько корней у уравнения $x^2 e^x = \pi$ и где они расположены.

Обозначим $f(x) := x^2 e^x - \pi$. Тогда $f'(x) = x(x+2)e^x$. Очевидно, $f'(x) = 0$ только при $x = 0$ и $x = -2$. Поскольку $f(x)$ и $f'(x)$ определены и непрерывны на всей числовой оси, точки –2 и 0 — единственные на Ox такие, в которых может происходить смена убывания функции $y = f(x)$ на возрастание или наоборот. Поэтому, найдя знаки значений (в том числе, бесконечных) $\lim\limits_{x\to-\infty} f(x)$, $f(-2)$, $f(0)$ и $\lim\limits_{x\to+\infty} f(x)$, т.е. заполнив таблицу знаков

$$f : \frac{-\infty}{-}\bigg|\frac{-2}{-}\bigg|\frac{0}{-}\bigg|\frac{+\infty}{+},$$

можно на основании теоремы 5.3 утверждать, что данное уравнение имеет

единственный корень, и этот корень положителен. Выяснив еще знаки $f(x)$ в точках $x = 1$ (минус) и $x = 2$ (плюс), область поиска корня данного уравнения с бесконечного промежутка $[0, +\infty)$ сужаем до промежутка единичной длины $[1, 2]$.

В ситуациях, далеких от рассмотренной идеальной, часто поступают следующим образом. Всю область определения (если она конечна) или какую-нибудь ее часть, вызывающую по тем или иным соображениям интерес, разбивают на отрезки точками x_i, расположенными на условно небольшом расстоянии h одна от другой (сюда включаются также граничные точки области определения). Вычислив значения $f(x)$ во всех этих точках (или только определив знаки $f(x_i)$), сравнивают их в соседних точках, т.е. проверяют, не выполнится ли на отрезке $[x_{i-1}, x_i]$ условие $f(x_{i-1})f(x_i) \le 0$. Если заранее известно количество корней в исследуемой области, то, измельчая шаг поиска h, таким процессом можно либо все их локализовать, либо довести процесс до состояния, позволяющего утверждать, что возможно наличие пар корней, не различимых с точностью $h = \varepsilon$. Этот хорошо приспособленный для вычислительных машин ***способ перебора*** является дорогим в смысле затрат на получение многочисленных пробных значений функции и не дает гарантий выявления количества и локализации всех корней в общем случае, что ограничивает сферу его применения.

Одна из проблем, в которую упирается решение задачи локализации корней, — это практическая невозможность точного вычисления значений функций. Она уже обсуждалась в главе 1. Вернемся к ней еще раз.

Из-за ограниченности разрядной сетки компьютера даже алгебраические функции вычисляются приближенно, вычисление же трансцендентных функций составляет тему отдельного разговора; ясно, что их значения за редкими исключениями по определению могут записываться лишь приближенно. Погрешности, с которыми вычисляются значения функции, порождают блуждания (иначе, флуктуации, случайные отклонения) ординат графика около средних значений. Если рассматривать небольшой участок графика какой-либо функции, построенной путем изображения всех ее точек, вычисленных на компьютере при дискретном изменении аргумента с очень мелким шагом h, с последующим соединением этих точек отрезками прямых, то окажет-

ся, что этот реальный график имеет пилообразную форму (рис. 5.2).

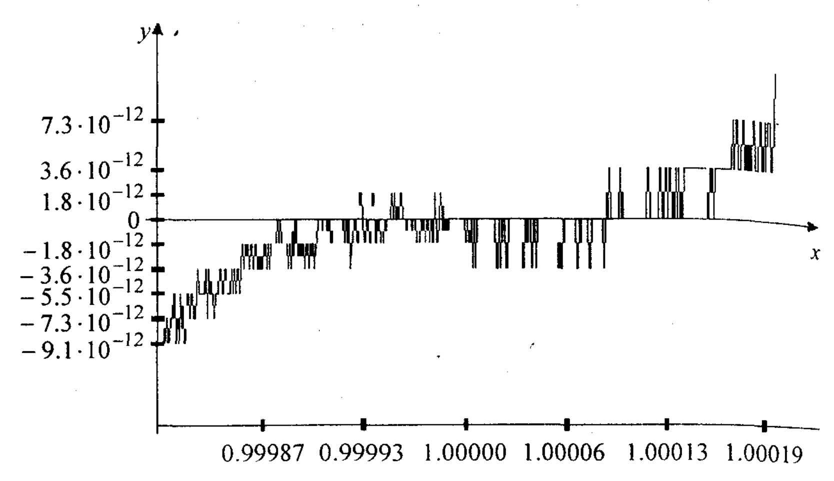


Рис. 5.2. Участок графика функции $y = x^3 - 3x^2 + 3x - 1$, *полученный на компьютере DX5-133 вычислением ее значений с точностью* $O\left(10^{-12}\right)$ *с шагом* $O\left(10^{-6}\right)$

Эти вычислительные погрешности ограничивают точность, с которой можно находить корень. В результате их влияния та или иная процедура сужения промежутка локализации корня рано или поздно приведет к такому промежутку, называемому ***промежутком*** (***отрезком***, ***интервалом***) ***неопределенности корня***, что любое число из этого промежутка с одинаковым успехом может быть принято за корень уравнения (сравните с промежутком (ξ − б.п.ξ, ξ + б.п.ξ), который фигурировал в § 1.6 при рассмотрении затронутой проблемы в несколько иной постановке).

Наличие флуктуаций значений функций, порождающих невозможность точного вычисления корней нелинейных уравнений, следует всегда иметь ввиду, применяя те или иные численные методы. Если не с самого начала итерационного процесса, то с приближением момента, когда погрешность метода становится сравнимой с погрешностью вычисления значений функций или выражений, входящих в расчетные формулы метода, целесообразно подключение приема Гарвика, упоминавшегося ранее (см. замечание 4.4).

5.2. МЕТОД ДИХОТОМИИ. МЕТОД ХОРД

Пусть функция $f(x)$ определена и непрерывна при всех $x \in [a,b]$ и на $[a,b]$ меняет знак, т.е. $f(a)f(b) < 0$. Тогда согласно теореме 5.1 уравнение (5.1) имеет на (a, b) хотя бы один корень. Возьмем произвольную точку $c \in (a, b)$. Будем называть в этом случае отрезок $[a,b]$ ***промежутком существования корня***, а точку c — ***пробной точкой***. Поскольку речь здесь идет лишь о вещественнозначных функциях вещественной переменной, то вычисление значения $f(c)$ приведет к какой-либо одной из следующих взаимоисключающих ситуаций:

а) $f(a)f(c)<0$; б) $f(c)f(b)<0$; в) $f(c)=0$.

Применительно к рассматриваемой задаче их можно интерпретировать так*):

а) корень находится на интервале (a, c);

б) корень находится на интервале (c, b);

в) точка c является искомым корнем.

Таким образом, одно вычисление значения функции позволяет уменьшить промежуток $[a,b]$ существования корня (ситуация а) или б)) или указать его значение (ситуация в), маловероятная в смысле «прямого попадания» пробной точкой c в корень, но вполне реальная в смысле выполнения приближенного равенства $f(c) \approx 0$, когда длина промежутка существования корня близка к длине промежутка его неопределенности). Ясно, что в зависимости от того, имеет место ситуация а) или б), описанная процедура одного шага сужения промежутка существования нуля непрерывной функции $f(x)$ может быть применена к промежутку $[a,c] \subset [a,b]$ или к $[c,b] \subset [a,b]$ соответственно и далее повторяться циклически. Такой простой и легко программируемый процесс называется ***методом дихотомии*** (от греческого слова, означающего деление на две части), ***методом бисекции***, ***методом вилки***, ***методом проб***. Если способ задания пробных точек c определен так, что последовательность длин получающихся в этом процессе промежутков существования корня стремится к нулю, то методом дихотомии можно найти какой-либо корень уравнения (5.1) с наперед заданной точностью.

Наиболее употребительным частным случаем метода дихотомии является ***метод половинного деления***, реализующий самый простой способ выбора пробной точки — деление проме-

*) Из-за допустимости неединственности корня в этой интерпретации уже нет взаимоисключаемости ситуаций.

жутка существования корня пополам. Выполнить приближенное вычисление с точностью ε корня уравнения (5.1) методом половинного деления при условии, что $f(x)$ непрерывна на $[a,b]$ и $f(a)f(b)<0$, можно, например, по следующей схеме:

Шаг 0. Задать концы отрезка a и b, функцию f, малое число $\varepsilon>0$ (допустимую абсолютную погрешность корня или полудлину его промежутка неопределенности), малое число $\delta>0$ (допуск, связанный с реальной точностью вычисления значений данной функции); вычислить (или ввести) $f(a)$.

Шаг 1. Вычислить $c:=0.5(a+b)$.

Шаг 2. Если $b-a<2\varepsilon$, положить $\xi:\approx c$ (ξ — корень) и остановиться.

Шаг 3. Вычислить $f(c)$.

Шаг 4. Если $|f(c)|<\delta$, положить $\xi:\approx c$ и остановиться.

Шаг 5. Если $f(a)f(c)<0$, положить $b:=c$ и вернуться к шагу 1; иначе положить $a:=c$, $f(a):=f(c)$ и вернуться к шагу 1.

Замечание 5.1. В упрощенных вариантах схем реализации метода половинного деления обходятся без введения допуска δ. В таком случае в шаге 4 вместо неравенства $|f(c)|<\delta$ используют равенство $f(c)=0$. Тогда разветвление алгоритма, диктуемое шагами 4 и 5, можно производить сравнением с нулем (с тремя исходами) произведения $f(a)f(c)$.

За один шаг метода половинного деления промежуток существования корня сокращается ровно вдвое. Поэтому, если за k-е приближение этим методом к корню ξ уравнения (5.1) примем точку x_k, являющуюся серединой полученного на k-м шаге отрезка $[a_k, b_k]$ в результате последовательного сужения данного отрезка $[a,b]$, полагая $a_1:=a$, $b_1:=b$, то придем к неравенству

$$|\xi-x_k|<\frac{b-a}{2^k} \qquad \forall k\in\mathbf{N} \tag{5.3}$$

(априори, ξ — любая точка интервала (a_k, b_k), и расстояние от нее до середины этого интервала не превосходит половины его длины. Это как раз и видим в (5.3) при $k=1$).

Неравенство (5.3), с одной стороны, позволяет утверждать, что последовательность (x_k) имеет предел — искомый корень ξ уравнения (5.1); с другой стороны, являясь априорной оценкой абсолютной погрешности приближенного равенства $x_k\approx\xi$, дает возможность подсчитать число шагов (итераций) метода половинного деления, достаточное для получения корня ξ с заданной

точностью ε, для чего нужно лишь найти наименьшее натуральное k, удовлетворяющее неравенству

$$\frac{b-a}{2^k}<\varepsilon.$$

Используемый в методе половинного деления способ фиксирования пробной точки можно охарактеризовать как пассивный, ибо он осуществляется по заранее жестко заданному плану и никак не учитывает вычисляемые на каждом шаге значения функции. Логично предположить, что в семействе методов дихотомии можно достичь несколько лучших результатов, если отрезок $[a,b]$ делить точкой c на части не пополам, а пропорционально величинам ординат $f(a)$ и $f(b)$ графика данной функции $f(x)$. Это означает, что точку c есть смысл находить как абсциссу точки пересечения оси Ox с прямой, проходящей через точки $A(a; f(a))$ и $B(b; f(b))$, иначе, с хордой AB дуги $A\xi B$ (рис. 5.3).

Запишем уравнение прямой, проходящей через две данные точки A и B:

$$\frac{y-f(a)}{f(b)-f(a)}=\frac{x-a}{b-a}.$$

Отсюда, полагая $y=0$ (уравнение оси Ox), $x=c$ (обозначение искомой точки пересечения прямой AB с осью Ox), находим

$$c=a-\frac{f(a)(b-a)}{f(b)-f(a)}. \tag{5.4}$$

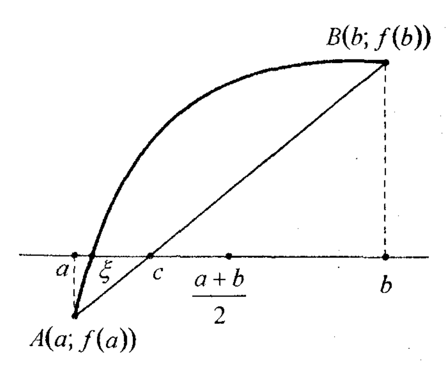


Рис. 5.3. Дуга графика функции $f(x)$ и стягивающая ее хорда

Метод, получающийся в развитие метода дихотомии таким фиксированием пробной точки, называют ***методом хорд***,

методом пропорциональных частей, ***методом линейной интерполяции***. Все названия метода вполне естественны и отражают различные подходы к его выводу или интерпретации. Иногда (в последнее время реже) используют еще и название ***правило ложного положения*** или ***regula falsi*** [3, 72, 140].

Существует несколько версий реализации метода хорд. Одна из них — подсчет значений c по формуле (5.4) в рамках алгоритма типа рассмотренного выше алгоритма половинного деления, где следует положить $f(b) := f(c)$ при $f(a)f(c) < 0$. Длина промежутка локализации корня при этом может не стремиться к нулю, поэтому обычно счет ведется до совпадения значений c на двух соседних итерациях с точностью ε (лучше, с точностью $\frac{m\varepsilon}{M-m}$, если $0 < m \le |f'(x)| \le M \quad \forall x \in [a,b]$, обоснование этого критерия см. в [61]).

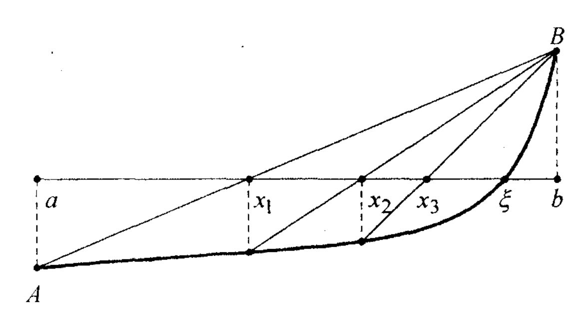


Рис. 5.4. Приближения к корню нелинейного уравнения по методу хорд

Так как для л и н е й н о й функции $f(x)$ метод хорд дает корень ξ точно за один шаг при любой длине отрезка $[a,b]$, то можно рассчитывать на его довольно быструю сходимость, если $f(x)$ близка к линейной. При определенных достаточно жестких условиях можно доказать соответствующие утверждения о монотонной сходимости, получить более точные оценки и упростить алгоритм (см., например, [61]). Однако в общем случае, если на функцию $f(x)$ не накладывать дополнительных ограничений, может оказаться, что метод хорд будет проигрывать в скорости методу половинного деления; чтобы убедиться в этом, достаточно взглянуть на рис. 5.4, демонстрирующий возможное поведение нескольких приближений по методу хорд.

5.3. ТИПЫ СХОДИМОСТЕЙ ИТЕРАЦИОННЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Чтобы более объективно судить о скорости сходимости тех или иных итерационных методов, введем следующие понятия [68 и др.].

Пусть некоторый итерационный процесс генерирует последовательность $(x_k)_{k=0}^{\infty}$, имеющую пределом x^*.

Определение 5.1. *Сходимость последовательности* (x_k) *к* x^* *называется* ***линейной*** (*соответственно, итерационный процесс —* ***линейно сходящимся***), *если существует такая постоянная* $C \in (0, 1)$ *и такой номер* k_0, *что*

$$\left|x^* - x_{k+1}\right| \le C\left|x^* - x_k\right| \quad \forall k \ge k_0, \tag{5.5}$$

и ***сверхлинейной***, *если существует такая положительная последовательность* $(C_k)_{k=0}^{\infty}$, *что* $C_k \underset{k\to\infty}{\to} 0$ *и*

$$\left|x^* - x_{k+1}\right| \le C_k\left|x^* - x_k\right| \quad \forall k \in \mathbf{N}_0. \tag{5.6}$$

Определение 5.2. *Говорят, что последовательность* (x_k) ***сходится*** *к* x^* ***по меньшей мере с p-м порядком*** (*соответственно,* ***итерационный процесс имеет по меньшей мере p-й порядок***), *если найдутся такие константы* $C > 0$ *и* $p \ge 1$, *что*

$$\left|x^* - x_{k+1}\right| \le C\left|x^* - x_k\right|^p \tag{5.7}$$

при всех $k \in \mathbf{N}_0$, *начиная с некоторого* $k = k_0$.

Фиксируя в определении 5.2 значение $p = 1$, видим, что линейно сходящийся процесс можно называть ***процессом первого порядка***; значению $p = 2$ в (5.7) соответствует ***квадратично сходящийся процесс***, $p = 3$ означает ***кубическую сходимость****).

К линейной сходимости применяют также термин ***сходимость со скоростью геометрической прогрессии***. Объяснение

*) В качестве известных примеров методов высоких порядков в пространстве $\mathbf{R}_n$ следует обратить внимание на изученное в § 3.6 семейство итерационных процессов обращения матриц.

ему можно найти в том, что определяющее линейную сходимость неравенство (5.5) между абсолютными погрешностями $(k+1)$-го и k-го приближений к предельной точке x^* означает существование последовательности положительных чисел ε_k, мажорирующих эти погрешности и связанных соотношением $\varepsilon_{k+1} = C\varepsilon_k$, т.е. являющихся членами геометрической прогрессии со знаменателем $C = \varepsilon_{k+1}/\varepsilon_k$ (= const). Отсюда следует также естественность в определении 5.1 условия $C<1$, чтобы последовательность погрешностей была убывающей, иначе и речи не может быть о сходимости (в определении 5.2 для предельного случая $p=1$ также следует ограничить C единицей; при $p>1$ в этом, вообще говоря, нет необходимости; проанализируйте, почему?).

Если требуемой в неравенстве (5.5) константы C не удается найти, но установлено неравенство (5.6) с $C_k \underset{k\to\infty}{\to} C \in (0,1)$, то в этом случае говорят об ***асимптотически линейной сходимости*** (пример такой сходимости дает степенной метод в § 4.4 или рассматриваемый далее в § 6.4 метод Бернулли). Аналогично можно определить ***асимптотически*** p***-й порядок***.

Имеются и другие понятия и термины, позволяющие более тонко классифицировать сходимость итерационных последовательностей и рассматривать ее как бы под разными углами зрения (см., например, [139, 140, 176, 178]). Как правило, они вводятся в более общем случае конечномерных или бесконечномерных нормированных пространств. Ничто не мешает и данные здесь определения распространить на многомерный случай, достаточно лишь элементы последовательности (x_k) и предельный элемент x^* считать n-мерными векторами $(n \ge 1)$ или матрицами, а вместо модуля использовать норму.

Среди нескольких способов охарактеризовать скорость сходимости итерационных последовательностей наиболее четко оформились два способа: опирающиеся на ***q-сходимость*** (от англ. *quotient* — частное) и на ***r-сходимость*** (от англ. *root* — корень). Происхождение этих терминов можно связать соответственно с признаками Даламбера (через отношение) и Коши (через арифметический корень), применяемыми для установления абсолютной сходимости ряда

$$x_0 + (x_1 - x_0) + (x_2 - x_1) + \ldots + (x_k - x_{k-1}) + \ldots, \tag{5.8}$$

что равносильно установлению сходимости данной последовательности (x_k), так как сходимость ряда (5.8) означает существование предела его частичных сумм

$$S_1 = x_0,\ \ S_2 = x_1,\ \ S_3 = x_2,\ \ \ldots,\ \ S_{k+1} = x_k,\ \ \ldots .$$

Четкие определения и различия между этими двумя типами сходимостей можно найти в [139]; более существенно эти различия проявляются в многомерном случае. Здесь же главное понимать, что представление о порядке сходимости того или иного метода важно для возможности сравнить его с другими; более точно для этого подходит знание порядка q-сходимости (что и определено выше), порядок же r-сходимости говорит лишь о наличии такой последовательности положительных чисел, которая, сходясь к нулю с этим порядком, мажорирует последовательность величин $\left|x^* - x_k\right|$. Не вникая в тонкости, можно сказать, что обычно, изучая итерационный метод, устанавливают факт сходимости итерационной последовательности (x_k) к искомому элементу x^* и получают апостериорные и априорные оценки погрешности, а о порядке метода (в том или ином смысле, не всегда уточняя, в каком) судят или на основе неравенства типа (5.7), или по априорной оценке погрешности вида

$$\left|x^* - x_k\right| \le C\nu^{p^k}, \tag{5.9}$$

где $C>0$, $\nu \in (0,1)$ — некоторые константы, а $p \ge 1$ — порядок метода, или по неравенству вида

$$\left|x_{k+1} - x_k\right| \le C\left|x_k - x_{k-1}\right|^p, \tag{5.10}$$

показывающему скорость сближения членов итерационной последовательности и являющемуся ключевым для установления сходимости и получения оценок погрешностей. Чаще всего, разные способы приводят к одному и тому же значению p, хотя это и не гарантировано.

Коснемся еще одного аспекта понятия сходимости итерационного метода. В приведенных выше определениях отождествлялись сходимость итерационного процесса и сходимость итерационной последовательности, порождаемой этим процессом; при этом негласно считалось, что последовательность (x_k) уже как бы фиксирована заданием начальной точки x_0. Больший же интерес представляет сходимость множества всевозможных итерационных последовательностей, генерируемых итерационным методом при варьировании x_0 в границах некоторой области. Итерационные методы, дающие в пределе решение данной задачи при любом начальном приближении x_0, называются ***глобально сходящимися***. Если же сходимость итерационной последовательности (x_k) к искомому элементу x^* имеет место лишь при задании x_0 из некоторой, вообще говоря, достаточно малой

окрестности x^*, то соответствующий итерационный метод называют ***локально сходящимся***.

Так, рассмотренные выше методы дихотомии можно отнести к глобально сходящимся методам, так как с их помощью всегда можно получить какой-нибудь из корней уравнения $f(x)=0$, если начать итерационный процесс с отрезка $[a,b]$ любой длины, входящего в область непрерывности $f(x)$, лишь бы было выполнено условие $f(a)f(b)<0$. При этом, как видно из оценки (5.3), имеющей форму (5.9), метод половинного деления нужно считать линейно сходящимся методом, т.е. он сходится со скоростью геометрической прогрессии со знаменателем $q=0.5$ и имеет [186] ***среднюю скорость сходимости*** $-\ln q=\ln 2$. Метод хорд также является методом первого порядка и в зависимости от свойств $f(x)$ может иметь как бо́льшую, так и меньшую, чем $\ln 2$, среднюю скорость (часто бо́льшую).

5.4. МЕТОД НЬЮТОНА

Одним из популярнейших итерационных методов решения нелинейных уравнений, что связано с его идейной простотой и быстрой сходимостью, является ***метод Ньютона****). Правило построения итерационной последовательности (x_k) здесь получают или из геометрических соображений, откуда другое название этого метода — ***метод касательных***, говорящее само за себя, или из аналитических путем подмены данной нелинейной функции ее линейной моделью на основе формулы конечных приращений Лагранжа или формулы Тейлора, в связи с чем метод Ньютона также называют ***методом линеаризации***. В любом случае, говорить о нахождении нуля функции $f(x)$ методом Ньютона можно лишь в предположении, что данная функция обладает достаточной гладкостью.

Для простоты будем считать, что функция $f(x)$ дважды дифференцируема на отрезке $[a,b]$, содержащем корень ξ уравнения (5.1).

Пусть $x_k \in [a,b]$ — уже известный член последовательности приближений к ξ, полученный конструируемым методом (или заданное начальное приближение x_0 при $k=0$). Для любого x из $[a,b]$ можно записать формальное представление $f(x)$ по

---

*) В зарубежной литературе его часто называют ***методом Ньютона–Рафсона*** [68, 115, 140, 176].

формуле Тейлора

$$f(x)=f(x_k)+f'(x_k)(x-x_k)+\frac{1}{2}f''(\Theta_k)(x-x_k)^2, \qquad (5.11)$$

где $\Theta_k \in [a,\ b]$ — некоторая точка между x и x_k.

Так как корень ξ — потенциально произвольная точка отрезка $[a,b]$, то разложение (5.11) справедливо и для $x=\xi$, т.е. существует точка $\Theta_k=\overline{\Theta_k}$ такая, что

$$f(\xi)=f(x_k)+f'(x_k)(\xi-x_k)+\frac{1}{2}f''(\overline{\Theta_k})(\xi-x_k)^2.$$

Но $f(\xi)=0$, и если точка $\overline{\Theta_k}$ известна, то корень ξ можно точно найти из квадратного уравнения

$$f(x_k)+f'(x_k)(\xi-x_k)+\frac{1}{2}f''(\overline{\Theta_k})(\xi-x_k)^2=0. \qquad (5.12)$$

Считая, что значение x_k близко к ξ, т.е. разность $\xi-x_k$ по модулю достаточно мала, можно рассчитывать, что величина $(\xi-x_k)^2$ будет тем более малой. На этом основании отбросим в (5.12) последнее слагаемое и подменим квадратное уравнение (5.12) линейным уравнением. Естественно, что при этом будет найден не корень ξ, а некоторая другая точка, которую обозначим x_{k+1}.

Таким образом, итерационный процесс Ньютона определяется линейным уравнением

$$f(x_k)+f'(x_k)(x_{k+1}-x_k)=0 \qquad (5.13)$$

или в явном виде формулой

$$x_{k+1}=x_k-\frac{f(x_k)}{f'(x_k)}, \qquad (5.14)$$

где $k=0,1,2,\ldots$ и предполагается, что по крайней мере на элементах последовательности (x_k) первая производная данной функции в нуль не обращается*).

Если в равенстве (5.13) фиксированную точку x_{k+1} заменить переменной x, а 0 в правой части — переменной y, то в

---

*) Как видим, процесс вычислений по методу Ньютона не требует знания второй производной. Можно обойтись без нее и при выводе (другим способом), но при этом усложняется изучение метода.

полученном легко узнать уравнение касательной к кривой $y = f(x)$, проведенной к ней в точке $(x_k;\ f(x_k))$. Отсюда **геометрический смысл метода Ньютона**: *приближения к корню ξ совершаются по абсциссам точек пересечения касательных к графику данной функции, проводимых в точках, соответствующих предыдущим приближениям* (рис. 5.5).

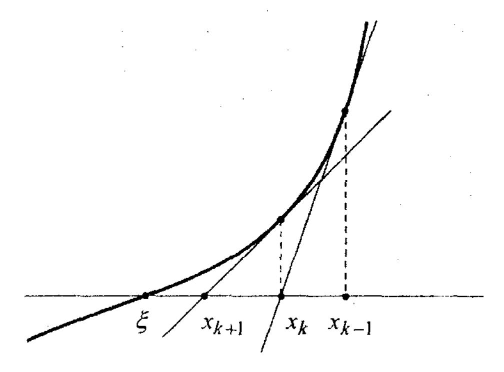


Рис. 5.5. Приближения к корню нелинейного уравнения методом касательных

Изучение сходимости метода Ньютона (осуществимость процесса вычисления элементов последовательности (x_k) по формуле (5.14) в пределах заданного отрезка, сходимость к корню ξ данного уравнения, порядок метода, оценки погрешности и критерии окончания процесса построения приближений, условия на выбор начального приближения) проводится при более ограничительных требованиях к данной функции $f(x)$.

Интуитивно ясно (из вида формулы (5.14), рассуждений при ее выводе и из ее геометрического смысла), что сходимость (x_k) к ξ будет тем быстрее и говорить о ней можно тем увереннее, чем ближе функция $f(x)$ к линейной и чем круче ее график пересекает ось абсцисс; так что есть смысл потребовать от $f(x)$, чтобы по модулю вторая ее производная была ограничена сверху, а первая — снизу. При этих условиях обратимся сначала к исследованию быстроты сходимости итерационного метода Ньютона (5.14) в предположении, что факт его осуществимости и сходимости к корню ξ сомнений не вызывает.

Теорема 5.4. *Пусть функция $f(x)$ удовлетворяет условиям*

$$(\text{A}) \qquad \begin{bmatrix} |f'(x)| \geq \alpha > 0 \\ |f''(x)| \leq \beta < \infty \end{bmatrix} \qquad \forall x \in [a, b].$$

Тогда, если члены последовательности (x_k), определяемые методом Ньютона (5.14), *при любом фиксированном $k \in \mathbf{N}_0$ принадлежат отрезку $[a, b]$ и эта последовательность сходится на $[a, b]$ к корню ξ уравнения* (5.1), *то справедливы неравенства* ($\forall k \in \mathbf{N}_0$):

$$|\xi - x_{k+1}| \leq \frac{\beta}{2\alpha}|\xi - x_k|^2, \qquad (5.15)$$

$$|\xi - x_{k+1}| \leq \frac{\beta}{2\alpha}|x_{k+1} - x_k|^2. \qquad (5.16)$$

(Первое из этих неравенств в соответствии с определением 5.2 позволяет считать (5.14) ***методом второго порядка***, а второе, являясь апостериорной оценкой погрешности, может служить в качестве критерия останова процесса вычислений).

Доказательство. Отметим, что требование принадлежности точек x_k отрезку $[a, b]$ дает право пользоваться оговоренным условиями (А) поведением данной функции в этих точках и их малых окрестностях.

Подставляя в правую часть формулы (5.12) вместо нуля левую часть равенства (5.13) (оба равенства истинны для рассматриваемых точек x_k, x_{k+1}), имеем:

$$f'(x_k)\xi - f'(x_k)x_k + \frac{1}{2}f''\left(\overline{\Theta_k}\right)(\xi - x_k)^2 = f'(x_k)x_{k+1} - f'(x_k)x_k.$$

Это равенство можно записать в виде точной связи между ошибками k-го и $(k+1)$-го приближений*):

$$\xi - x_{k+1} = -\frac{f''\left(\overline{\Theta_k}\right)}{2f'(x_k)}(\xi - x_k)^2, \qquad (5.17)$$

из которой, переходя к модулям и привлекая условия (А), получаем первое из доказываемых неравенств.

*) Для элемента x_{k+1} последовательности (x_k) приближений к корню ξ уравнения $f(x) = 0$ величину $\xi - x_{k+1}$ называют ***ошибкой***, $x_{k+1} - x_k$ — ***поправкой***, а $f(x_{k+1})$ — ***невязкой***.

Для доказательства второго неравенства сначала установим связь между невязкой $(k+1)$-го приближения и разностью соседних (k-го и $(k+1)$-го) приближений. С этой целью в формулу Тейлора (5.11) подставим $x = x_{k+1}$. Имеем

$$f(x_{k+1}) = f(x_k) + f'(x_k)(x_{k+1} - x_k) + \frac{1}{2} f''(\widetilde{\Theta}_k)(x_{k+1} - x_k)^2,$$

и воспользуемся тем, что согласно (5.13), первые два слагаемых правой части равенства в совокупности дают нуль. Таким образом,

$$f(x_{k+1}) = \frac{1}{2} f''(\widetilde{\Theta}_k)(x_{k+1} - x_k)^2$$

и значит,

$$|f(x_{k+1})| \le \frac{\beta}{2} |x_{k+1} - x_k|^2. \qquad (5.18)$$

Применим теперь формулу Лагранжа к разности $f(\xi) - f(x_{k+1})$. Согласно этой формуле, между точками ξ и x_{k+1} найдется точка τ_{k+1} такая, что

$$f(\xi) - f(x_{k+1}) = f'(\tau_{k+1})(\xi - x_{k+1}).$$

Принимая во внимание, что $f(\xi) = 0$, а также условия (A), получаем неравенство между абсолютными величинами невязок и ошибок приближений:

$$|\xi - x_{k+1}| \le \frac{1}{\alpha} |f(x_{k+1})|.$$

Усилив его с помощью неравенства (5.18), приходим к доказываемой оценке (5.16).

Теорема полностью доказана.

Замечание 5.2. Последнее неравенство, справедливое для приближений x_{k+1} к нулю ξ дифференцируемой функции $f(x)$ независимо от способа их получения, обосновывает контроль точности по невязкам, а именно, оправдывает критерий окончания итерационного процесса

$$|f(x_k)| < \varepsilon \Rightarrow \xi := x_k \text{ с точностью } \varepsilon,$$

если в окрестности ξ выполняется требование $|f'(x)| > 1$, и говорит о необходимости учитывать множитель $\frac{1}{\alpha}$, т.е. добиваться выполнения неравенства

$$|f(x_k)| < \alpha\varepsilon, \quad \text{если} \quad \alpha < 1.$$

Чтобы выяснить, при каких условиях на выбор начального приближения x_0 начинающаяся с него последовательность (x_k), генерируемая методом Ньютона (5.14), будет сходиться к корню ξ, проитерируем формально неравенство (5.15). Имеем:

при $k = 0$ $\quad |\xi - x_1| \le \frac{\beta}{2\alpha} |\xi - x_0|^2;$

при $k = 1$ $\quad |\xi - x_2| \le \frac{\beta}{2\alpha} |\xi - x_1|^2 \le \frac{\beta}{2\alpha} \left(\frac{\beta}{2\alpha}\right)^2 |\xi - x_0|^{2^2};$

при $k = 2$ $\quad |\xi - x_3| \le \frac{\beta}{2\alpha} |\xi - x_2|^2 \le \frac{\beta}{2\alpha} \left(\frac{\beta}{2\alpha}\right)^2 \left(\frac{\beta}{2\alpha}\right)^{2^2} |\xi - x_0|^{2^3};$

далее по индукции получаем

$$|\xi - x_k| \le \left(\frac{\beta}{2\alpha}\right)^{1+2+2^2+\ldots+2^{k-1}} |\xi - x_0|^{2^k} =$$

$$= \left(\frac{\beta}{2\alpha}\right)^{\frac{2^k - 1}{2-1}} |\xi - x_0|^{2^k} = \frac{2\alpha}{\beta} \left(\frac{\beta}{2\alpha}\right)^{2^k} |\xi - x_0|^{2^k},$$

т.е. при любых $k \in N$

$$|\xi - x_k| \le \frac{2\alpha}{\beta} \left(\frac{\beta}{2\alpha} |\xi - x_0|\right)^{2^k}. \qquad (5.19)$$

Отсюда следует, что

$$x_k \to \xi, \quad \text{если} \quad \frac{\beta}{2\alpha} |\xi - x_0| < 1.$$

Таким образом, появилась возможность судить о том, насколько далеко от корня ξ можно брать начальное приближение x_0 в зависимости от свойств данной функции $f(x)$, и по априорной оценке (5.19) заранее подсчитывать число итераций, достаточное для вычисления корня с заданной точностью, если есть оценка близости x_0 к ξ.

Теорема 5.5. *Пусть для функции* $f(x)$ *на отрезке* $[a, b]$ *выполнены условия* (A).

Тогда, если интервал $J := \left(\xi - \frac{2\alpha}{\beta}, \xi + \frac{2\alpha}{\beta}\right)$ *содержит-*

ся в $[a,b]$, то при произвольном выборе x_0 из J для определяемой методом Ньютона (5.14) *последовательности* (x_k):

1) $x_k \in J \quad \forall k \in \mathbf{N}$;

2) $\exists \lim\limits_{k\to\infty} x_k = \xi$ *и* $f(\xi)=0$;

3) *справедливо утверждение теоремы* 5.4 *и оценка* (5.19).

Доказательство. Прежде всего заметим, что условие $x_0 \in J$ равносильно неравенству

$$\nu := \frac{\beta}{2\alpha}|\xi - x_0| < 1, \tag{5.20}$$

к которому мы пришли выше, анализируя неравенство (5.19) (рис. 5.6).

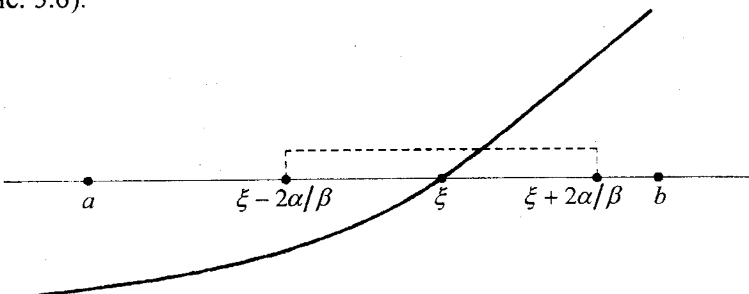


Рис. 5.6. Промежуток выбора начальной точки в методе Ньютона

Покажем осуществимость процесса (5.14) в интервале J.

Так как $x_0 \in J \subset [a,b]$, то, согласно условиям (A), $f'(x_0) \neq 0$, и значение x_1 по формуле (5.14) может быть получено. Более того, справедливо неравенство (5.15) при $k=0$. Следовательно,

$$|\xi - x_1| \le \frac{\beta}{2\alpha}|\xi - x_0|^2 = \frac{2\alpha}{\beta}\nu^2 < \frac{2\alpha}{\beta},$$

т.е. $x_1 \in J$. Аналогично из индукционного предположения, что $x_k \in J$ при некотором $k \in \mathbf{N}$, т.е. что $|\xi - x_k| < \dfrac{2\alpha}{\beta}$, используя (5.15), получаем:

$$|\xi - x_{k+1}| \le \frac{\beta}{2\alpha}|\xi - x_k|^2 < \frac{\beta}{2\alpha}\cdot\left(\frac{2\alpha}{\beta}\right)^2 = \frac{2\alpha}{\beta} \;\Rightarrow\; x_{k+1} \in J.$$

Итак, согласно принципу математической индукции, все элементы последовательности (x_k) лежат в $J \subset [a,b]$, и значит, можно воспользоваться неравенством (5.19), из которого тут же следует, что $\xi = \lim x_k$, и заключение предыдущей теоремы.

То, что точка ξ — середина заявленного в теореме интервала J — есть корень уравнения (5.1), если заранее это неизвестно, устанавливается переходом к пределу в формуле (5.14) (с учетом условия $f'(x) \neq 0 \;\forall x \in [a,b]$).

Слабым местом только что доказанной теоремы 5.5 является то, что точкой отсчета при построении промежутка применимости метода Ньютона служит корень ξ, который как раз и неизвестен. Более естественно за центр такого промежутка принимать некоторую конкретную точку x_0 — начальное приближение, полагая по тем или иным соображениям (например, геометрическим), что в ее окрестности должен быть корень. Имеется ряд теорем подобного типа (см. [99, 102, 129, 140, 158] и др.). Как правило, им присуща некоторая громоздкость и трудная проверяемость условий.

Заменив условия (A) на требование знакопостоянства первой и второй производных данной функции, означающих монотонность и определенную выпуклость ее графика, докажем простую теорему несколько иного плана.

Теорема 5.6 [61]. *Пусть на отрезке $[a,b]$ функция $f(x)$ имеет первую и вторую производные постоянного знака и пусть*

$$f(a)f(b) < 0.$$

Тогда, если точка x_0 выбрана на $[a,b]$ так, что

$$f(x_0)f''(x_0) > 0, \tag{5.21}$$

то начатая с нее последовательность (x_k), определяемая методом Ньютона (5.14), *монотонно сходится к корню $\xi \in (a,b)$ уравнения* (5.1).

Доказательство опирается на теорему Вейерштрасса о сходимости монотонной ограниченной последовательности.

Положим, для определенности, что

$$f'(x) > 0 \quad \text{и} \quad f''(x) > 0 \quad \forall x \in [a,b].$$

Тогда $f(a)<0$, $f(b)>0$, и в качестве начальной точки x_0, удовлетворяющей условию (5.21),*) можно взять любую точку из промежутка $(\xi, b]$ (наличие корня ξ, единственного в (a, b), условиями теоремы обеспечено), и при этом x_0 из $(\xi, b]$ будет $f(x_0)>0$.

Покажем ограниченность последовательности (x_k) снизу и ее монотонное убывание. Из равенства (5.12) при $k=0$, т.е. из

$$f(x_0)+f'(x_0)(\xi-x_0)+\frac{1}{2}f''(\Theta_0)(\xi-x_0)^2=0,$$

в силу положительности последнего слагаемого, следует, что

$$f(x_0)+f'(x_0)(\xi-x_0)<0.$$

Отсюда, учитывая положительность $f'(x)$, имеем

$$\xi<x_0-\frac{f(x_0)}{f'(x_0)}.$$

Это неравенство показывает, что

$$\xi<x_1=x_0-\frac{f(x_0)}{f'(x_0)}<x_0$$

(значит, $f(x_1)>0$). По индукции устанавливаем, что

$$\xi<x_{k+1}<x_k \quad \forall\, k\in\mathbf{N}_0,$$

т.е. последовательность (x_k) монотонно убывает и ограничена снизу самим корнем ξ, следовательно, имеет предел. Сходимость (x_k) именно к корню, как и в предыдущей теореме, получаем переходом к пределу в равенстве (5.14).

Остальные комбинации знаков производных рассматриваются аналогично. Теорема доказана.

Замечание 5.3. Нарушение условия Фурье (5.21) на выбор начального приближения x_0 при выполненных требованиях к знакопостоянству производных может отразиться лишь на первом приближении: его перебросит через корень ξ на другую часть отрезка $[a, b]$ (рис. 5.7). Если при этом x_1 не окажется за пределами отрезка $[a, b]$, то далее итерационный

---

*) Называемому иногда *условием* ***Фурье***.

процесс Ньютона пойдет монотонно (сформулируйте самостоятельно аналог теоремы 5.6 без условия **Фурье**).

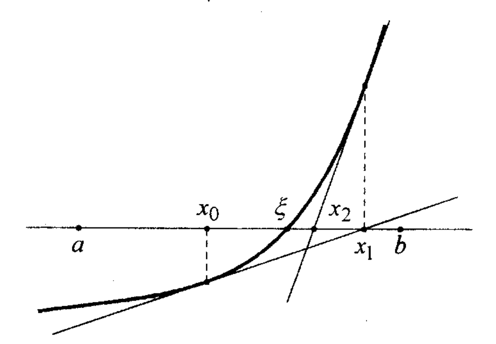


Рис. 5.7. Поведение первых приближений по методу Ньютона при нарушении условия Фурье на выбор x_0

Обратимся вновь к зафиксированной теоремами 5.4 и 5.5 квадратичной сходимости метода Ньютона, которая имеет место и в условиях теоремы 5.6. О поведении такого процесса можно получить наглядное представление из оценки (5.19), схематично записанной в виде

$$|\xi-x_k|\le C\cdot\nu^{2^k},$$

где $\nu\in(0,1)$ (см. (5.20)), а $C\left(=\frac{2\alpha}{\beta}\right)$ — некоторая положительная постоянная. Допустим, что функция $f(x)$ и приближение x_0 к корню ξ таковы, что $C=1$, а $\nu=0.1$, т.е. абсолютные погрешности убывают по закону $|\xi-x_k|\le 0.1^{2^k}$. Подставляя сюда $k=1,2,3,\ldots$, имеем: $|\xi-x_1|\le 10^{-2}$, $|\xi-x_2|\le 10^{-4}$, $|\xi-x_3|\le 10^{-8}$, $|\xi-x_4|\le 10^{-16}$ и т.д.

Как видим, *квадратично сходящийся процесс в идеале*, т.е. если он реализуется точно, *должен давать удвоение числа верных знаков на каждой итерации, начиная с некоторой*. Такой высокий темп установления верных цифр искомого корня не только позволяет получить корень с большой точностью сравнительно небольшим количеством вычислений, но и обеспечивает хорошую численную устойчивость метода, а также меньшую критичность к правилу окончания итерационного процесса (часто здесь применяют упрощенное правило останова

$$|x_k-x_{k-1}|\le\varepsilon \;\Rightarrow\; \xi:\approx x_k).$$

5.5. ПРИМЕНЕНИЕ МЕТОДА НЬЮТОНА К ВЫЧИСЛЕНИЮ ЗНАЧЕНИЙ ФУНКЦИЙ

Элементарные функции, чаще всего, вычисляются с помощью приближения их подходящими многочленами (об этом еще пойдет речь далее, см. гл. 9). В некоторых же случаях для этих целей применяют итерационные методы, в частности, базирующиеся на методе Ньютона.

Пусть требуется найти значение заданной функции φ в заданной точке a. Считая a произвольной точкой из области $D(\varphi)$ или какой-либо ее подобласти, функциональное соответствие

$$x = \varphi(a)$$

зададим неявно уравнением

$$F(a,\ x) = 0 \tag{5.22}$$

таким, чтобы: 1) оно было локально эквивалентным (в окрестности точки a) данному; 2) функция F была дифференцируема по второму аргументу; 3) функции F и F' были легко вычислимы.

При каждом фиксированном a уравнение (5.22) можно считать уравнением типа (5.1) и получать приближенно его корень — требуемое значение $x = \varphi(a)$ — методом Ньютона (5.14). Для уравнения (5.22) формула (5.14) принимает вид

$$x_{k+1} = x_k - \frac{F(a, x_k)}{F'_x(a, x_k)}, \tag{5.23}$$

где $k = 0, 1, 2, \ldots$, а x_0 — задаваемое начальное приближение к $\varphi(a)$.

В качестве более конкретного примера применения такого подхода выведем из формулы (5.23) ***правило Ньютона вычисления арифметических корней.***

Пусть a — данное положительное число, а $n \geq 2$ — данный натуральный показатель корня.

Очевидна связь между задачей вычисления вещественного значения

$$x = \sqrt[n]{a}$$

и задачей нахождения положительного корня уравнения

$$x^n - a = 0.$$

Приняв $F(a,x) := x^n - a$, находим $F'_x(a,x) = nx^{n-1}$, и согласно (5.23), процесс приближений к $\sqrt[n]{a}$ определяем формулой

$$x_{k+1} = x_k - \frac{x_k^n - a}{nx_k^{n-1}}$$

или в другом виде

$$x_{k+1} = \frac{1}{n}\left[(n-1)x_k + \frac{a}{x_k^{n-1}}\right],$$

где $k = 0, 1, 2, \ldots$, а $x_0 > 0$ задается.

Еще из глубокой древности известен частный случай полученного правила Ньютона — ***процесс Герона***

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{a}{x_k}\right),$$

применяемый для извлечения квадратных корней. Здесь $F(a, x) = x^2 - a$, $F'(a, x) = 2x$, $F''(a, x) = 2$, т.е. при $x > 0$ выполняются предварительные условия теоремы 5.6, и для монотонной сходимости (x_k) к $\sqrt{a}$ достаточно лишь удовлетворить условию Фурье (5.21), т.е. взять x_0 таким, чтобы было $x_0^2 > a$. Легко убедиться, что это условие будет выполнено, если взять $x_0 = 2^{0.5m}$, если m — четное, и $x_0 = 2^{0.5(m+1)}$, если m — нечетное, где $m \in \mathbf{Z}$ такое, что $a = 2^m q$ и $q \in [0.5, 1]$ (a «зажимается» между двумя соседними целыми степенями двойки: $2^{m-1} \leq a < 2^m$)*).

Другим примером использования метода Ньютона в форме (5.23) может служить вычисление обратной величины данного числа a, иначе говоря, выполнение операции деления с помощью других арифметических операций.

Аналогично предыдущему зададим искомое $x = \frac{1}{a}$ как корень уравнения

$$a - \frac{1}{x} = 0.$$

Подставляя $F(a, x) = a - \frac{1}{x}$ и $F'_x(a, x) = \frac{1}{x^2}$ в (5.23), после упро-

---

*) Не заботясь о монотонности последовательности приближений, обычно ограничиваются заданием $x_0 = 2^{[0.5m]}$.

щения получаем итерационный процесс без делений*)

$$x_{k+1} = x_k(2 - ax_k), \quad k = 0, 1, 2, \ldots. \tag{5.24}$$

Здесь также считаем a и x положительными, и очевидно, что для любых $x > 0$

$$F' > 0, \quad \text{а} \quad F'' = -\frac{2}{x^3} < 0,$$

т.е. условия теоремы 5.6 будут выполнены, если взять $x_0 \in \left(0, \frac{1}{a}\right)$.

Непосредственное изучение процесса (5.24) показывает, что область выбора начального приближения x_0 может быть в два раза расширена (правда монотонность приближений при этом уже не гарантируется).

Действительно, рассмотрим связь между поправками $(k+1)$-го и k-го приближений:

$$\frac{1}{a} - x_{k+1} = \frac{1}{a} - 2x_k + ax_k^2 = a\left(\frac{1}{a} - x_k\right)^2$$

(равенство типа (5.7), подтверждающее, что (5.24) — процесс второго порядка). Отсюда последовательным итерированием приходим к равенству

$$\frac{1}{a} - x_k = \frac{1}{a}(1 - ax_0)^{2^k},$$

из которого следует, что сходимость (x_k) к $\frac{1}{a}$ имеет место в случае, когда $|1 - ax_0| < 1$, т.е. при $x_0 \in \left(0, \frac{2}{a}\right)$.

Обычно за начальное приближение, удовлетворяющее условию $x_0 \in \left(0, \frac{1}{a}\right)$, берут число 2^{-m}, где $m \in \mathbf{Z}$ то же, что и в предыдущем примере.

---

*) Легко увидеть аналогию между этим процессом и рассмотренным ранее процессом Шульца второго порядка для обращения матриц (сравните (5.24) с (3.34) при $m = 1$).

5.6. МОДИФИКАЦИИ МЕТОДА НЬЮТОНА. МЕТОД СЕКУЩИХ

Вновь обратимся к теоремам 5.4–5.6 о сходимости метода Ньютона. Их условия предполагают неравенство нулю производной данной функции $f(x)$ на промежутке $[a, b]$, где применяется метод. А это означает, что они регламентируют применение метода Ньютона только для нахождения простых нулей функции $f(x)$, поскольку для кратного корня ξ уравнения (5.1) имеет место равенство $f'(\xi) = 0$. Действительно, пусть ξ — m-кратный корень $(m \geq 2)$; тогда функция $f(x)$ представима в виде*)

$$f(x) = (x - \xi)^m f_1(x)$$

и ее производная $f'(x) = (x - \xi)^{m-1} \cdot [mf_1(x) + (x - \xi)f_1'(x)]$ обращается в нуль при $x = \xi$.

Согласно (5.14), формально нужно, чтобы производная не равнялась нулю в точках x_k последовательности приближений, в предельной же точке ξ допустимо обращение производной в нуль. Как показывает пример 5.3, итерационный процесс Ньютона может сходиться и в этом случае, т.е. когда ξ является кратным корнем уравнения (5.1), но сходимость при этом — только линейная.

Если заведомо известно число m — показатель кратности корня ξ, то для ускорения сходимости метода Ньютона в формулу (5.14) рекомендуется ввести корректирующий множитель m:

$$x_{k+1} = x_k - m \cdot \frac{f(x_k)}{f'(x_k)}. \tag{5.25}$$

Такую модификацию будем называть ***методом Ньютона–Шрёдера****). Доказательство сверхлинейной сходимости этого метода можно найти в [158], где он называется ***методом Ньютона с параметром***, а также в [176] и в [140], где имеется ссылка на содержащую этот метод работу Э.Шрёдера.

---

*) Это представление здесь принимается за определение m-й кратности корня ξ. Часто ξ считают корнем кратности m, если $f(\xi) = f'(\xi) = \ldots = f^{(m-1)}(\xi) = 0$, а $f^{(m)}(\xi) \neq 0$ (см., например, [158]).

) В [72] первое введение параметра m в формулу (5.14) приписывается Е.Бодевигу (1949 г.), в связи с чем (5.25) там называют ***методом Ньютона–Бодевига*.

Пример 5.3. Для функции $f(x)=(x-1)^2$ корень $\xi=1$ — двукратный. Подстановка этой функции $f(x)$ и ее производной $f'(x)=2(x-1)$ в формулу (5.14) определяет процесс

$$x_{k+1}=x_k-\frac{x_k-1}{2}, \qquad (5.26)$$

или проще,

$$x_{k+1}=\frac{1}{2}(x_k+1).$$

Если начать его с $x_0=2$, то на каждой последующей итерации будем получать все более близкие к $\xi=1$ значения:

$$x_1=1.5, \quad x_2=1.25, \quad x_3=1.125, \quad x_4=1.0625, \quad \ldots$$

Вычитая 1 из обеих частей (5.26), приходим к равенству

$$x_{k+1}-1=\frac{1}{2}(x_k-1),$$

означающему, что сходимость последовательности (x_k) к 1 — точно линейная. В то же время, введение множителя $m=2$ в (5.26) в соответствии с (5.25) приводит к стационарной последовательности $x_{k+1}=1 \quad \forall k\in \mathbf{N}_0$.

Не следует думать, что из (5.25) всегда будет получаться $x_{k+1}=\xi$. Рассмотрим менее утрированный пример.

Пример 5.4. Функция $f(x)=x(x-1)^2$ с тем же двухкратным корнем $\xi=1$ с помощью формул (5.14) и (5.25) порождает следующие процессы:

метод Ньютона	*метод Ньютона–Шрёдера*
$x_{k+1}=\dfrac{2x_k^2}{3x_k-1}$;	$x_{k+1}=\dfrac{x_k(x_k+1)}{3x_k-1}$;
$x_0=2,$	$x_0=2,$
$x_1=1.6,$	$x_1=1.2,$
$x_2=1.347368,$	$x_2=1.000116,$
$x_3=1.193517,$	$x_3=1.000000.$
...	...

Налицо эффективность коррекции метода Ньютона введением в него показателя кратности корня.

Цель всех последующих видоизменений основной формулы (5.14) метода Ньютона — уменьшение вычислительных затрат, связанных с необходимостью вычисления производной на каждом итерационном шаге.

Самый простой выход на этом пути — использование на каждом шаге одного и того же шагового множителя $\dfrac{1}{f'(x_0)}$, т.е. счет по формуле

$$x_{k+1}=x_k-\frac{f(x_k)}{f'(x_0)}, \quad k=0,1,2,\ldots\,. \qquad (5.27)$$

Такой метод называют ***модифицированным*** или ***упрощенным методом Ньютона***[*]. Он имеет очевидную геометрическую интерпретацию: в начальной точке x_0 проводится касательная к графику $y=f(x)$ (первый шаг основного и модифицированного методов Ньютона совпадают), а во всех последующих точках $x_1, x_2, \ldots$ проводятся прямые, параллельные этой касательной (рис. 5.8).

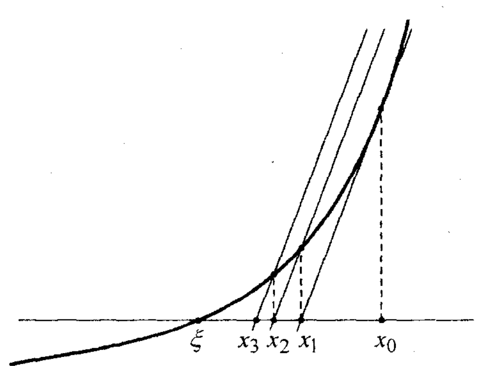


Рис. 5.8. Приближения к корню нелинейного уравнения модифицированным (упрощенным) методом Ньютона

При такой модификации метод Ньютона утрачивает высокую скорость сходимости (процесс не реагирует на изменение наклона кривой к оси абсцисс при приближении к корню) и вместо квадратичной имеет лишь скорость сходимости геометрической прогрессии, что станет очевидным несколько позже (см. § 6.1).

На получение сверхлинейной скорости сходимости при видоизменении метода Ньютона (5.14) можно надеяться в случае, когда $f'(x_k)$ при каждом $k\in\mathbf{N}$ подменяется не одним и тем же числом $f'(x_0)$, а некоторым близким к $f'(x_k)$ значением, которое может быть найдено (при каждом k свое) через значения

---

[*] Иногда его называют еще ***огрубленный метод Ньютона*** [158].

данной функции. Для таких аппроксимаций*) $f'(x_k)$ можно использовать, например, определение производной. Имеем:

$$f'(x_k)=\lim_{h\to 0}\frac{f(x_k+h)-f(x_k)}{h},$$

и при малых h (произвольного знака) получаем приближенное равенство

$$f'(x_k)\approx\frac{f(x_k+h)-f(x_k)}{h}, \qquad (5.28)$$

позволяющее производную приближенно подменять так называемым ***разностным отношением*** (подробнее об этом см. гл. 13). Подстановка (5.28) в (5.14) приводит к итерационной формуле

$$x_{k+1}=x_k-\frac{f(x_k)\cdot h}{f(x_k+h)-f(x_k)}, \qquad (5.29)$$

где k=0, 1, 2,..., а h — малый параметр, которым должен распорядиться вычислитель.

Ясно, что при каждом k в формуле (5.28) может быть свое значение h, т.е. в формуле (5.29) вместо постоянного параметра h имеет смысл использовать связанный с номером итерации параметр h_k, т.е. вести вычисления по формуле

$$x_{k+1}=x_k-\frac{f(x_k)h_k}{f(x_k+h_k)-f(x_k)}. \qquad (5.30)$$

Итерационный метод, определяемый формулами (5.29) или (5.30), назовем ***разностным методом Ньютона****).

Так как равенство (5.28) можно сделать сколь угодно точным за счет малости шага h разностного отношения (теоретически; практически это далеко не так из-за потерь точности при вычитании близких чисел), то по непрерывности можно утверждать асимптотически квадратичную скорость сходимости разностного метода Ньютона при определенных условиях.

Рассмотрим соображения, которыми следует руководствоваться при задании последовательности параметров h_k в разностном методе Ньютона (5.30). В любом случае, будем исходить из постулата, что с ростом k значения $|h_k|$ должны убывать, чтобы при приближении x_k к корню ξ производная $f'(x_k)$ все более точно аппроксимировалась разностным отношением $[f(x_k+h_k)-f(x_k)]/h_k$.

Первое, что можно здесь предложить, так это задать какое-либо значение h_0, а каждое последующее значение параметра получать рекуррентным равенством $h_{k+1}=\delta h_k$, где $\delta\in(0,1)$ – некоторое фиксированное число. Например, можно положить $h_0=0.1$, $h_1=0.01$, $h_2=0.001$ и т.д. Очевиден недостаток такого подхода — отсутствие связи между скоростью сходимости (x_k) к ξ и скоростью убывания $|h_k|$ (может оказаться, что x_k еще не имеет достаточной близости к ξ, а значение $|h_k|$ настолько мало, что значения $f(x_k+h_k)$ и $f(x_k)$ реально не различимы; противоположная ситуация чревата большой потерей скорости сходимости или, еще хуже, нарушением канонического развития итерационного процесса).

Если учесть, что при зафиксированных в теореме 5.4 условиях (А) $f(x_k)\to 0$ с той же скоростью, что и $x_k\to\xi$ (см. неравенство (5.18)), есть смысл полагать в (5.30) $h_k:=f(x_k)$. Разумеется, это можно делать на той стадии итерационного процесса, когда значения $|f(x_k)|$ уже достаточно малы (иначе теряет силу (5.28)). При таких h_k формула (5.30) принимает вид

$$x_{k+1}=x_k-\frac{(f(x_k))^2}{f(x_k+f(x_k))-f(x_k)} \qquad (5.31)$$

и называется ***методом Стеффенсена***. Подчеркнем еще раз его сугубо локальный характер сходимости, но зато *сходимость* эта *квадратичная* [140, 158].

При приблизительно равных затратах на вычисление значений данной функции и ее производной ни один из рассмотренных выше вариантов разностного метода Ньютона не дает выигрыша по сравнению с основным методом (5.14), поскольку каждый из них требует два вычисления функции на каждом итерационном шаге, не увеличивая при этом скорость сходимости. Построим такую модификацию метода Ньютона, развивая далее его разностный аналог (5.30), в которой на один шаг итерации приходилось бы только одно вычисление функции.

Опираясь на то, что необходимым условием сходимости некоторой последовательности x_k к пределу ξ, как это следует из (5.8), является сходимость к нулю последовательности

---

*) *Approximare* (лат.) — приближаться.

) Другие названия — ***конечноразностный* [68] и ***дискретный*** [139] ***метод Ньютона***. Сходимость этого метода изучается в [68]; там же можно найти рекомендации по сопряжению шага дискретизации h с точностью машинных вычислений.

разностей $x_k - x_{k-1}$ (причем с той же скоростью, см. (5.10)), положим в (5.30)

$$h_k := x_{k-1} - x_k, \quad \text{откуда} \quad x_{k-1} = x_k + h_k.$$

В результате этого из (5.30) получаем итерационный процесс

$$x_{k+1} = x_k - \frac{f(x_k)(x_{k-1} - x_k)}{f(x_{k-1}) - f(x_k)}, \tag{5.32}$$

где $k = 1, 2, 3, \ldots$, а x_0 и x_1 должны задаваться.

Формула (5.32) определяет новый метод как ***двухшаговый*** (результат $(k+1)$-го шага зависит от результатов k-го и $(k-1)$-го шагов) и на каждой итерации требует вычисления только одного значения функции, другое же значение, фигурирующее в этой формуле, передается с предыдущего шага. Сравнив (5.32) с формулой (5.4), полученной из геометрических соображений, легко понять, что x_{k+1} есть абсцисса точки пересечения с осью Ox прямой, проведенной через точки $(x_{k-1};\ f(x_{k-1}))$ и $(x_k;\ f(x_k))$, т.е. секущей (рис. 5.9). Отсюда название этого метода — ***метод секущих****).

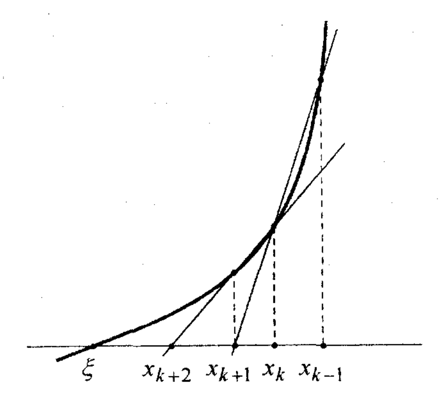


Рис. 5.9. Приближения к корню методом секущих

Можно сказать, что метод секущих и метод хорд определяются совершенно однотипными формулами, но порождающие их идеологии различны, что сказывается на свойствах и скорости

*) В [12] методом секущих называют метод хорд (с фиксированным концом).

сходимости генерируемых ими последовательностей приближений.

Выясним, по какому закону убывают погрешности приближений, получаемых методом секущих (5.32) в условиях (А) теоремы 5.4.

Вычитая равенство (5.32) из равенства $\xi = \xi$ и применяя в знаменателе формулу Лагранжа, имеем:

$$\xi - x_{k+1} = \xi - x_k + \frac{f(x_k)(x_{k-1} - x_k)}{f(x_{k-1}) - f(x_k)} = \xi - x_k + \frac{f(x_k)}{f'(\mu_k)}, \tag{5.33}$$

где μ_k — некоторая точка между приближениями x_k и x_{k-1}. Далее воспользуемся формулой Тейлора, согласно которой можно записать

$$f(x) = f(\xi) + f'(\xi)(x - \xi) + \frac{1}{2} f''(\theta)(x - \xi)^2,$$

откуда при $x = x_k$, $\theta = \theta_k$ с учетом того, что ξ — корень уравнения (5.1), следует представление

$$f(x_k) = f'(\xi)(x_k - \xi) + \frac{1}{2} f''(\theta_k)(x_k - \xi)^2.$$

Подставим это в (5.33) и вынесем в правой части общий множитель $\xi - x_k$:

$$\xi - x_{k+1} = \frac{\xi - x_k}{f'(\mu_k)} \cdot \left[f'(\mu_k) - f'(\xi) + \frac{1}{2} f''(\theta_k)(\xi - x_k) \right].$$

Вновь применяя формулу Лагранжа — теперь к разности производных — и переходя к модулям, получаем неравенство

$$|\xi - x_{k+1}| \le \frac{|\xi - x_k|}{|f'(\mu_k)|} \cdot \left[|f''(\nu_k)| \cdot |\xi - \mu_k| + \frac{1}{2} |f''(\theta_k)| |\xi - x_k| \right].$$

Так как для сходящейся к ξ нестационарной последовательности (x_k) справедливы неравенства

$$|\xi - \mu_k| < |\xi - x_{k-1}|, \quad |\xi - x_k| < |\xi - x_{k-1}|,$$

то в итоге можно записать искомую связь погрешностей в виде

$$|\xi - x_{k+1}| < \frac{3\beta}{2\alpha} |\xi - x_k| \cdot |\xi - x_{k-1}|, \tag{5.34}$$

где α и β — константы из условий (А).

Ясно, что неравенство (5.34) характеризует *метод секущих*

как *сверхлинейно сходящийся процесс*. Конкретный порядок метода секущих устанавливается следующим образом*).

Обозначим для краткости

$$\varepsilon_k := |\xi - x_k|, \quad C := \frac{3\beta}{2\alpha}$$

и, используя записанное в этих обозначениях неравенство (5.34)

$$\varepsilon_{k+1} < C\varepsilon_k \varepsilon_{k-1},$$

получим последовательно несколько первых оценок ε_k через степени ε_0 (полагая по определению $\varepsilon_1 < \varepsilon_0$). Имеем:

$$\text{при } k=1 \qquad \varepsilon_2 < C\varepsilon_1\varepsilon_0 < \frac{1}{C}(C\varepsilon_0)^2;$$

$$\text{при } k=2 \qquad \varepsilon_3 < C\varepsilon_2\varepsilon_1 < \frac{1}{C}(C\varepsilon_0)^3;$$

$$\text{при } k=3 \qquad \varepsilon_4 < C\varepsilon_3\varepsilon_2 < \frac{1}{C}(C\varepsilon_0)^5;$$

$$\text{при } k=4 \qquad \varepsilon_5 < C\varepsilon_4\varepsilon_3 < \frac{1}{C}(C\varepsilon_0)^8;$$

$$\text{при } k=5 \qquad \varepsilon_6 < C\varepsilon_5\varepsilon_4 < \frac{1}{C}(C\varepsilon_0)^{13} \quad \text{и т.д.}$$

Обратив внимание на то, что показатели в правых частях этих неравенств подчиняются закону «каждый последующий есть сумма двух предыдущих», нетрудно доказать в общем виде, что

$$\varepsilon_k < \frac{1}{C}(C\varepsilon_0)^{\Phi_k}, \qquad (5.35)$$

где (Φ_k) — последовательность ***чисел Фибоначчи****), определяемая рекуррентным соотношением

$$\Phi_{k+1} = \Phi_{k-1} + \Phi_k, \quad k = 1, 2, 3, \ldots, \quad \Phi_0 = \Phi_1 = 1.$$

---

*) Близкое к данному изложение можно найти в [140].

**) Фибона́ччи — псевдоним знаменитого итальянского математика Леонардо Пизанского (1180–1240 гг.).

Для общего члена Φ_k этой последовательности известна ***формула Бинэ***:

$$\Phi_k = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^{k+1} - \left(-\frac{1+\sqrt{5}}{2}\right)^{-(k+1)}\right].$$

Поскольку с ростом k роль второго члена в выражении Φ_k становится ничтожной, для достаточно больших значений k на основе (5.35) можно считать справедливым асимптотическое неравенство

$$|\xi - x_k| < C_1 \cdot \nu^{\left(\frac{1+\sqrt{5}}{2}\right)^k},$$

где C_1 и ν — некоторые новые постоянные (которые легко выписать). Полученная оценка вида (5.9) позволяет утверждать, что справедлива следующая теорема.

Теорема 5.7. *Метод секущих имеет порядок по крайней мере* $\frac{1+\sqrt{5}}{2}$ (≈ 1.618).

Замечание 5.4. Для многошаговых итерационных методов иногда вводят специфическое понятие порядка сходимости. Согласно [68], *j-шаговый итерационный метод, генерирующий сходящуюся к ξ последовательность (x_k), называется **j-шагово сходящимся с порядком p**, если* $|\xi - x_{k+j}| \le C|\xi - x_k|^p$. Так как из неравенства (5.34) после его усиления получается

$$|\xi - x_{k+1}| < \frac{3\beta}{2\alpha}|\xi - x_{k-1}|^2,$$

то можно сказать, что *метод секущих сходится двухшагово квадратично.*

Высокий порядок скорости сходимости метода секущих в сочетании с минимальными вычислительными затратами — одно вычисление значения функции на один итерационный шаг — выводит этот метод на первое место по эффективности решения скалярных уравнений вида (5.1) среди прочих итерационных методов. Это подтверждается как теоретическими, так и практическими наблюдениями*).

---

*) При сравнении с методом Ньютона и другими методами, использующими производные, в [140] предлагается приравнивать работу по вычислению значений функций и ее производных. Единица такой работы в [140] называется ***горнером***, а в [176] — ***единицей объема информационного запроса***.

8 Б-223

При применении метода секущих возникают вопросы, связанные с началом итерационного процесса и с его окончанием. Поскольку касательная к кривой есть предельное положение секущей, выбор начальной точки x_0 в методе секущих нужно осуществлять по тому же принципу, что и в методе касательных, например, привлекая условие Фурье (5.21); вторая же из начальных точек x_1, требуемая в двухшаговом методе (5.32), может быть взята в непосредственной близости от x_0 (понятие близости здесь, разумеется, условно), желательно между точкой x_0 и искомым корнем ξ.

Окончание счета по методу секущих, учитывая его быструю сходимость, можно контролировать с помощью проверок на малость модулей невязок, т.е. $|f(x_k)|$, или поправок $|x_k - x_{k-1}|$. Однако главное здесь — это суметь вовремя остановить процесс вычислений, не дожидаясь момента, когда погрешности вычислений начнут превосходить погрешность метода вследствие вычитания приближенно вычисляемых близких значений $f(x_{k-1})$ и $f(x_k)$ в знаменателе расчетной формулы (5.32). В этом плане, т.е. в численной устойчивости, метод секущих уступает методу Ньютона.

Если почти все рассмотренные выше методы можно отнести к классу методов линеаризации, имея ввиду, что в их основе лежит подмена исходной нелинейной модели линейной, построенной тем или иным способом, то следующим шагом должно быть построение классов методов «параболизации». Квадратичная модель (парабола) может быть получена, например, по формуле Тейлора или квадратичной интерполяцией (гл. 8), но в любом случае соответствующие итерационные формулы, хотя и дают более быстро сходящиеся последовательности приближений к корню, либо содержат старшие производные данной функции, либо являются слишком громоздкими и сложными как для исследования, так и для их применения.

Более важно обратить внимание на локальную сходимость таких простых и достаточно быстро сходящихся методов, как метод Ньютона и метод секущих, условия которой не так часто удается обеспечить. В связи с этим встает задача построения ***гибридных алгоритмов*** на базе двух или нескольких методов, соединяя быструю сходимость одних с глобальной сходимостью других. Принципы комбинирования методов в таких алгоритмах могут быть различными: можно «стартовать» с глобально сходящегося «медленного» метода и подключить быстросходящийся метод на финише для уточнения значения корня, а можно сразу начать процесс вычислений «быстрым» методом, но проводить корректировку получаемых им значений, пользуясь глобально сходящимся методом. Последний подход порождает,

например, следующий простейший гибридный алгоритм.

Метод Ньютона–метод половинного деления

Шаг 0. Задать начальное приближение x_0, положить $k := 0$.

Шаг 1. Вычислить $\tilde{x}_k = x_k - \dfrac{f(x_k)}{f'(x_k)}$.

Шаг 2. Если $|f(\tilde{x}_k)| < |f(x_k)|$, то $x_{k+1} := \tilde{x}_k$; иначе $\tilde{x}_k := \frac{1}{2}(x_k + \tilde{x}_k)$ и возвратиться к началу шага 2.

Шаг 3. Проверить на останов (работа алгоритма либо прекращается с $\xi :\approx x_{k+1}$, либо продолжается переходом к шагу 1 с $k := k+1$.

Приведенный алгоритм учитывает, что метод Ньютона вырабатывает локально правильное направление (убывания функции), но продвижение в этом направлении может оказаться чрезмерным, что и корректируется с помощью деления отрезка пополам, если не выполняется ***условие релаксации*** $|f(\tilde{x}_k)| < |f(x_k)|$ в шаге 2 (рис. 5.10).

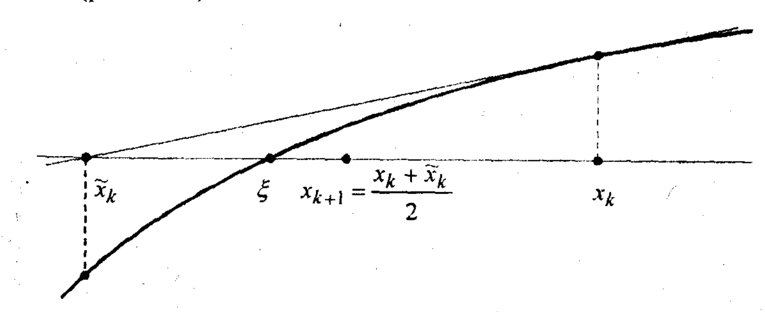


Рис. 5.10. Иллюстрация одного шага гибридного метода Ньютона–половинного деления

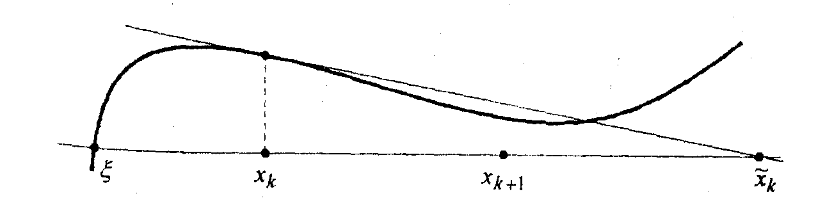


Рис. 5.11. Пример поведения гибридного метода Ньютона–половинного деления, когда x_{k+1} дальше от корня ξ, чем x_k

Такой гибрид трудно считать глобально сходящимся (рассмотрите, например, его дальнейшее поведение в ситуации, изображенной на рис. 5.11), но он позволяет расширить границы применимости метода Ньютона и хоть как-то вести процесс поиска корня в условиях неопределенности знаков производных. Разумеется, этот алгоритм весьма схематичен и требует некоторых усилий на его деталировку. Особенно важно решить, как в тех или иных случаях выполнить шаг 3. Большую роль здесь играет правильное сопряжение задаваемой точности решения задачи с погрешностью метода (иначе, с остаточной погрешностью) и с точностью выполнения арифметических операций на используемой вычислительной машине (вычислительной погрешностью), а также с точностью вычисления значений функций, что, вообще говоря, не одно и то же. Важные сведения о таких критериях окончания процессов поиска корней с привязкой их к реальным компьютерам можно почерпнуть в книге [68].

5.7. ПОЛЮСНЫЕ МЕТОДЫ НЬЮТОНА И СЕКУЩИХ

Построим новую модификацию метода Ньютона (5.14) решения уравнений вида (5.1), введя в него два параметра. В основу вывода этой модификации положим геометрические представления.

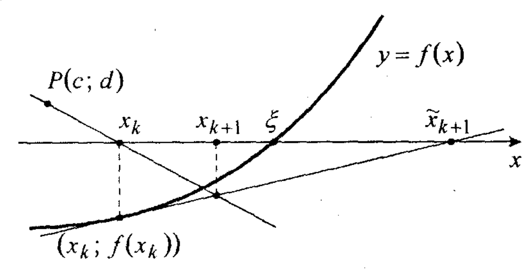


Рис. 5.12. К построению полюсного метода Ньютона

Возьмем на плоскости Oxy некоторую точку $P(c;d)$ (назовем ее ***полюсом***) и через нее и определяемую предыдущим приближением x_k точку $(x_k;0)$ проведем прямую (см. рис. 5.12). Новым приближением x_{k+1} к корню ξ уравнения $f(x)=0$ будем считать абсциссу точки пересечения этой прямой с касательной к графику функции $y=f(x)$, проведенной в точке $(x_k;f(x_k))$. Составив уравнения указанных прямых

$$y=\frac{d(x-x_k)}{c-x_k} \quad \text{и} \quad y=f'(x_k)(x-x_k)+f(x_k),$$

разрешаем получающееся отсюда приравниванием ординат уравнение

$$\frac{d(x-x_k)}{c-x_k}=f'(x_k)(x-x_k)+f(x_k)$$

относительно абсциссы x и найденное таким образом ее значение называем x_{k+1}. В результате приходим к итерационной формуле

$$x_{k+1}=x_k-\frac{f(x_k)}{f'(x_k)-\dfrac{d}{c-x_k}}, \quad k=0,1,2,\dots. \qquad (5.36)$$

Эта формула определяет *двухпараметрический одношаговый метод*, который в дальнейшем будем называть ***полюсным методом Ньютона****).

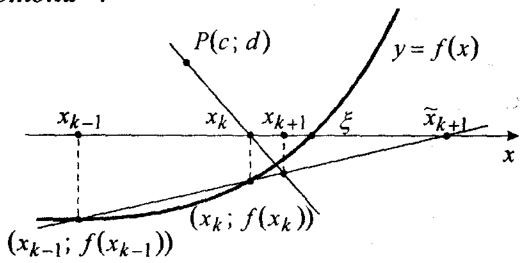


Рис. 5.13. К построению полюсного метода секущих

Аналогично, проводя через точки $(x_{k-1};f(x_{k-1}))$ и $(x_k;f(x_k))$ секущую

$$y=f(x_k)+\frac{f(x_{k-1})-f(x_k)}{x_{k-1}-x_k}(x-x_k)$$

(рис. 5.13), находим ее точку пересечения с прямой, проведенной через полюс $P(c;d)$ и $(x_k;0)$. Выражение абсциссы этой точки пересечения двух прямых задает *двухпараметрический двухшаговый итерационный процесс*

$$x_{k+1}=x_k-\frac{f(x_k)}{\dfrac{f(x_{k-1})-f(x_k)}{x_{k-1}-x_k}-\dfrac{d}{c-x_k}}, \quad k=1,2,\dots, \qquad (5.37)$$

*) Идея этого метода предложена и опробована П.В.Вержбицким (1989 г.).

который будем называть ***полюсным методом секущих***.

Легко видеть, что при равенстве нулю выражения $\frac{d}{c-x_k}$ (т.е. при $d=0$) новые методы (5.36) и (5.37) совпадают с базовыми для них классическими методами Ньютона (5.14) и секущих (5.32) соответственно, следовательно обобщают их, и что формула (5.37) может быть получена из формулы (5.36) аппроксимацией производной подобно тому, как формула секущих (5.32) была получена из формулы касательных (5.14).

Анализ рисунков 5.12 и 5.13, на которых через $\tilde{x}_{k+1}$ обозначены приближения по базовым методам (5.14) и (5.32) соответственно, показывает, что за счет удачного расположения полюса P полюсными модификациями методов Ньютона и секущих можно получить лучшее уточнение приближенного значения корня, чем посредством базового метода (сравните $|\xi - x_{k+1}|$ с $|\xi - \tilde{x}_{k+1}|$ на том и на другом рисунках). Нетрудно также представить графически ситуацию, когда полюсные варианты будут генерировать сходящиеся к корню последовательности, в то время как, например, ньютоновское приближение отбрасывается в бесконечность.

Изучим поведение погрешности k-го приближения, получаемого полюсным методом Ньютона (5.36), в предположении о существовании в некоторой окрестности корня ξ уравнения (5.1) (содержащей начальную точку x_0) непрерывной второй производной функции $f(x)$.

Запишем определяющую полюсный метод Ньютона формулу (5.36) в виде равенства

$$f(x_k)+\left(f'(x_k)-\frac{d}{c-x_k}\right)(x_{k+1}-x_k)=0 \qquad (5.38)$$

и воспользуемся тем, что в соответствии с формулой Тейлора

$$0=f(\xi)=f(x_k)+f'(x_k)(\xi-x_k)+\frac{1}{2}f''(\theta_k)(\xi-x_k)^2, \qquad (5.39)$$

где θ_k — некоторая точка между x_k и ξ. Приравнивая к левой части равенства (5.38) правую часть равенства (5.39), после упрощений получаем следующую связь между ошибками $(k+1)$-го и k-го приближений:

$$\xi - x_{k+1} = -\frac{f''(\theta_k)}{2f'(x_k)}(\xi-x_k)^2 - \frac{d\cdot(x_{k+1}-x_k)}{f'(x_k)(c-x_k)}. \qquad (5.40)$$

Лишь второе слагаемое в представлении (5.40) ошибки $(k+1)$-го приближения зависит от параметров c и d, выбор которых априори может как ухудшить, так и улучшить типично ньютоновскую связь ошибок $(k+1)$-го и k-го приближений, определяемую первым слагаемым (сравните с (5.17)).

Выразив поправку $x_{k+1}-x_k$ из формулы (5.36) и подставив ее в равенство (5.40), приходим к другой его форме:

$$\xi - x_{k+1} = -\frac{f''(\theta_k)}{2f'(x_k)}(\xi-x_k)^2 - \frac{d\cdot f(x_k)}{f'(x_k)[d-f'(x_k)(c-x_k)]}. \qquad (5.41)$$

Так как $f(x_k)$ с помощью линеаризации в окрестности корня ξ можно представить в виде

$$f(x_k)=f(\xi)+f'(\tau_k)(x_k-\xi)=f'(\tau_k)(x_k-\xi) \qquad (5.42)$$

(при некотором значении τ_k, расположенным между точками x_k и ξ), то учет этого в равенстве (5.41) позволяет рассчитывать, по меньшей мере, на линейный закон убывания погрешностей метода (5.36) при любых значениях параметров c и d, если только знаменатель второй дроби в (5.41) не стремится к нулю при $k\to\infty$.

Линейная связь (5.42) между невязкой $f(x_k)$ и ошибкой $\xi - x_k$ приближения x_k к корню ξ, рассматриваемая применительно к равенству (5.41), говорит о том, что если ординату d полюса P изменять пропорционально изменению значения функции $f(x)$ в текущей точке, то можно рассчитывать на квадратичную сходимость метода (5.36). Действительно, полагая, например, $d:=f(x_k)$, из (5.41) с учетом (5.42) получаем

$$\xi - x_{k+1} = -\left[\frac{f''(\theta_k)}{2f'(x_k)}+\frac{(f'(\tau_k))^2}{f'(x_k)[f(x_k)-f'(x_k)(c-x_k)]}\right](\xi-x_k)^2, \qquad (5.43)$$

что показывает допустимость квадратичной сходимости последовательности (x_k).

При ограниченной снизу абсолютной величине производной функции $f(x)$ множитель при $(\xi-x_k)^2$ в равенстве (5.43) все же может оказаться большим, если будет слишком малой по модулю величина

$$u_k := f(x_k)-f'(x_k)(c-x_k).$$

Однако за счет такого выбора другого параметра (т.е. c), при котором он изменялся бы согласованно с изменением f и f', величина u_k может быть сделана фиксированной. Например, полагая

$$c := x_k + \frac{f(x_k)-2}{f'(x_k)},$$

при любом $k \in \mathbf{N}_0$ будем иметь значение $u_k = 2$. При таком выборе d и c из равенства (5.43) следует равенство

$$\xi - x_{k+1} = -\frac{f''(\theta_k) + (f'(\tau_k))^2}{2f'(x_k)}(\xi - x_k)^2, \qquad (5.44)$$

означающее квадратичную связь ошибок $(k+1)$-го и k-го приближений. К этому же равенству придем и в случае фиксирования

$$d := -f(x_k), \quad c := x_k - \frac{f(x_k)+2}{f'(x_k)},$$

однако, как легко увидеть из рис. 5.12, достаточно ограничиться выбором $d = f(x_k)$, так как при $d = -f(x_k)$ и соответствующем значении c получаем симметричный относительно точки $(x_k; 0)$ полюс, с которым приходим к тому же значению очередного приближения x_{k+1}.

Итогом проведенных рассуждений служит следующая теорема.

Теорема 5.8. *Пусть корень ξ уравнения* (5.1), *начальная точка x_0 и элементы последовательности (x_k), полученной* ***методом Ньютона с подвижным полюсом***

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k) - \dfrac{d_k}{c_k - x_k}}, \quad k = 0, 1, 2, \ldots \qquad (5.45)$$

*принадлежит отрезку $[a, b]$, на котором выполняются неравенства**)

$$0 < \alpha \le |f'(x)| \le \gamma, \quad |f''(x)| \le \beta.$$

Тогда при

$$d_k = f(x_k), \quad c_k = x_k + \frac{f(x_k)-2}{f'(x_k)} \qquad (5.46)$$

быстрота сходимости последовательности (x_k) характеризуется неравенством

$$|\xi - x_{k+1}| \le \frac{\beta + \gamma^2}{2\alpha}|\xi - x_k|^2. \qquad (5.47)$$

---

*) Сравните с условиями (А) теоремы 5.4.

Из представленного теоремой 5.8 результата вытекает следующее утверждение.

Следствие 5.1. *Пусть для метода Ньютона с подвижным полюсом* (5.45), (5.46) *в условиях теоремы* 5.8 *начальное приближение x_0 к корню ξ уравнения* (5.1) *выбрано так, что*

$$t := \frac{\beta + \gamma^2}{2\alpha}|\xi - x_0| < 1. \qquad (5.48)$$

Тогда имеет место сходимость (x_k) к ξ с априорной оценкой погрешности

$$|\xi - x_k| \le \frac{2\alpha}{\beta + \gamma^2} t^{2^k} \quad \forall\, k \in \mathbf{N}. \qquad (5.49)$$

Действительно, положив $q := \dfrac{\beta + \gamma^2}{2\alpha}$, итерированием неравенства (5.47) получаем:

$$|\xi - x_k| \le q|\xi - x_{k-1}|^2 \le q(q|\xi - x_{k-2}|^2)^2 =$$
$$= q^{1+2}|\xi - x_{k-2}|^{2^2} \le q^{1+2+2^2}|\xi - x_{k-3}|^{2^3} \le \ldots$$
$$\ldots \le q^{1+2+2^2+\ldots 2^{k-1}}|\xi - x_0|^{2^k} = \frac{1}{q}(q|\xi - x_0|)^{2^k},$$

т.е.

$$|\xi - x_k| \le \frac{2\alpha}{\beta + \gamma^2}\left(\frac{\beta + \gamma^2}{2\alpha}|\xi - x_0|\right)^{2^k}.$$

Последнее неравенство показывает, что требование (5.48) является достаточным для сходимости x_k к ξ, и справедливость оценки (5.49).

Замечание 5.5. Анализируя равенство (5.44), подмечаем, что при задаваемом формулами (5.46) способе фиксирования координат подвижного полюса $P_k(c_k; d_k)$ скорость сходимости последовательности приближений, генерируемых процессом (5.45), будет выше для функций $f(x)$, которые в окрестности корня ξ выпуклы вверх. Можно рассчитывать, что сходимость (x_k) будет сверхквадратичной, если при этом вторая производная будет пропорциональна квадрату первой производной, т. е. для семейства функций $f(x, p, c_1, c_2)$, удовлетворяющих дифференциальному уравнению

$$f''(x) = -p(f'(x))^2, \quad p > 0.$$

Легко убедиться, что решением последнего является семейство функций

$$f(x) = \frac{1}{p} \ln (c_1 px + c_2), \tag{5.50}$$

где c_1, c_2 — произвольные постоянные, при которых данная функция имеет смысл, а $p > 0$ — коэффициент пропорциональности второй производной и квадрата первой, каким-то образом влияющий на быстроту сходимости. Ясно, что для функций вида (5.50) оценка (5.47) (и, тем более, вытекающая из нее оценка (5.49)) будет слишком грубой.

Выбор значения 2 фигурирующей в выражении параметра c_k постоянной обусловлен лишь целью выровнять коэффициенты для более простого вида связи погрешностей соседних итерационных шагов при переходе от равенства (5.43) к равенству (5.44) и, соответственно, оценок (5.47), (5.49). Вместо числа 2 можно ввести априори любое другое действительное число; будем считать его параметром ν. Смысл его в том, что это фиксируемое значение должно приниматься выражением $d - f'(x_k)(c - x_k)$ (см. (5.41)), т.е. в случае подвижного полюса должно быть

$$d_k - f'(x_k)(c_k - x_k) = \nu. \tag{5.51}$$

Отсюда имеем

$$c_k = x_k + \frac{d_k - \nu}{f'(x_k)}$$

или, в соответствии с изменением другой координаты подвижного полюса $P_k(c_k\,;\, d_k)$,

$$d_k = f(x_k), \quad c_k = x_k + \frac{f(x_k) - \nu}{f'(x_k)}. \tag{5.52}$$

Подставив (5.52) и (5.51) в равенство (5.41), с учетом (5.42) получаем равенство

$$\xi - x_{k+1} = -\frac{1}{f'(x_k)} \left[\frac{f''(\theta_k)}{2} + \frac{(f'(\tau_k))^2}{\nu} \right] (\xi - x_k)^2, \tag{5.53}$$

из которого видно, что для ускорения сходимости при задании координат подвижного полюса по формулам (5.52) знак параметра $\nu\ (\neq 0)$ желательно брать противоположным знаку второй производной данной функции (к сожалению, это не улучшит вытекающую из (5.53) оценку

$$|\xi - x_{k+1}| \le \frac{\nu\beta + 2\gamma^2}{2\nu\alpha} |\xi - x_k|^2).$$

Равенство (5.53) также показывает, что при правильном согласовании знаков и том или ином определенном взаимном поведении первой и второй производных данной функции $f(x)$ в окрестности корня ξ должны найтись такие значения ν, при которых однопараметрический метод (5.45), (5.52) покажет заведомо более высокую скорость сходимости, чем метод Ньютона (5.14) (кстати, отметим, что характеризующее быстроту сходимости ньютоновского процесса равенство (5.17) получается из (5.53) при $\nu \to \infty$).

Итак, одна из версий полюсного метода Ньютона состоит в следующем (назовем это ***однопараметрическим полюсным методом Ньютона***).

Предположим, что функция $f(x)$ в уравнении (5.1) дважды дифференцируема и в окрестности искомого корня, содержащей начальное приближение x_0, первая и вторая производные сохраняют знак. Возьмем некоторую постоянную $\nu\ (\neq 0)$ такую, чтобы

$$\nu f''(x_0) < 0$$

(если нет определенных соображений по ее выбору, полагаем

$$\nu = -2 \operatorname{sign} f''(x_0)). \tag{5.54}$$

При $k = 0, 1, 2, \ldots$ вычисляем

$$d_k = f(x_k), \quad x_{k+1} = x_k - \frac{d_k(\nu - d_k)}{\nu f'(x_k)} \tag{5.55}$$

пока не выполнится какой-либо критерий окончания итерационного процесса (например, пока не окажется $|d_k|$ меньше заданного малого числа $\varepsilon > 0$).

Эффект, который можно получить, применяя полюсный метод Ньютона (5.55) вместо классического метода Ньютона (5.14), продемонстрируем на следующем простом примере.

Пример 5.5. Начиная с $x_0 = \frac{1}{e}$, будем вычислять приближения к корню $\xi = 1$ уравнения $\ln x = 0$ методом Ньютона (5.14) и по формулам (5.55), в которых в соответствии с (5.54) зафиксируем $\nu = 2$. Результаты этих вычислений, а именно, сами приближения x_k и их невязки $f(x_k)$ на каждой итерации k представлены табл. 5.1.

Таблица 5.1

Сравнительное поведение приближений к корню $\xi = 1$ уравнения $\ln x = 0$, получаемых классическим и однопараметрическим полюсным методами Ньютона

k	Классический метод Ньютона (5.14)		Полюсный метод Ньютона (5.55) с $v = 2$	
	x_k	$f(x_k)$	x_k	$d_k = f(x_k)$
0	0.36787944117	–1.00000000000	0.36787944117	–1.00000000000
1	0.73575888234	–0.30685281944	0.91969860293	–0.08370926813
2	0.96152856982	–0.03923100060	0.99990817502	–0.00009182920
3	0.99925029771	–0.00074998345	1.00000000000	–0.00000000000
4	0.99999971890	–0.00000028110		
5	1.00000000000	0.00000000000		

Наблюдаемая в табл. 5.1. скорость «размножения нулей» перед первой значащей цифрой невязки в методе (5.55) подтверждает высказанные выше соображения в пользу возможной сверхквадратичной сходимости метода Ньютона с подвижным полюсом. Геометрическая иллюстрация первых двух шагов метода (5.55), соответствующих приведенным в табл. 5.1 числовым результатам, показана на рис. 5.14.

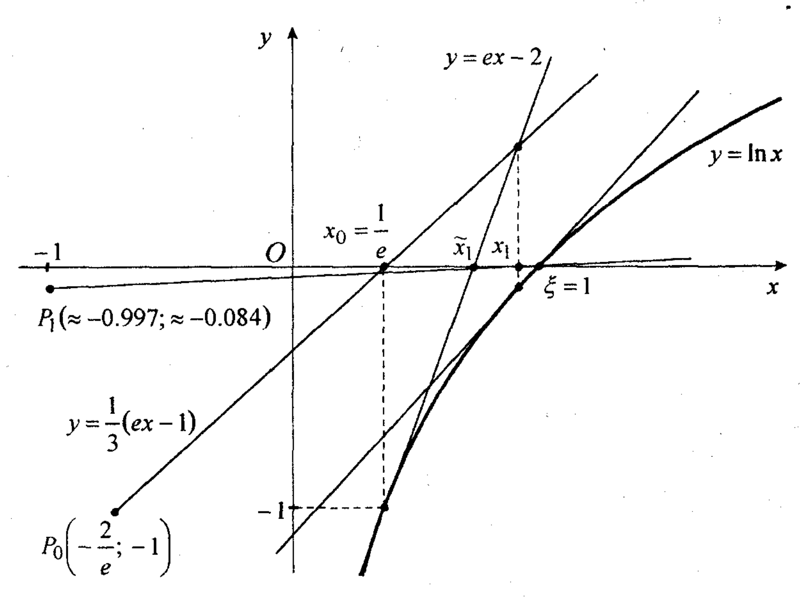


Рис. 5.14. *Два шага метода* (5.55) *при* $v = 2$
($\tilde{x}_1$ — *первое ньютоновское приближение*, x_2 *не различимо с* ξ)

Практически такое же поведение процесса (5.55) имеет место и в случае выбора более «плохой» начальной точки, $x_0 = e$ (для которой условие Фурье (5.21) не выполняется). При том же значении параметра $v = 2$ по формулам (5.55) находим

$$x_1 = 0.99361162491, \qquad f(x_1) = -0.00640886808,$$
$$x_2 = 0.99999312529, \qquad f(x_2) = -0.00000687474,$$
$$x_3 = 0.99999999999, \qquad f(x_3) = -0.00000000001,$$

т.е. с высокой точностью имеем $x_3 \approx \xi$, в то время как по методу Ньютона на первом шаге получается приближение $x_1 = e - \frac{\ln e}{1/e} = 0$, в котором функция $\ln x$ не определена, и процесс вычислений сразу застопоривается.

Не следует думать, что существенный выигрыш полюсной модификации метода Ньютона обнаруживается лишь на функциях, близких по поведению к семейству (5.50) (к которому, кстати, принадлежит использованная в примере 5.5 функция $y = \ln x$). Примеров, в которых есть смысл применить полюсный метод Ньютона вместо классического, великое множество (см. хотя бы упр. 5.8). Однако имеются примеры и противоположного толка. Так, в случае линейной функции $y = x$ обычный метод Ньютона даст точное решение за один шаг при любом начальном приближении x_0, чего не добиться никакими ухищрениями, применяя нетривиальный полюсный метод Ньютона (т.е. когда параметр d в нем отличен от нуля). Отсюда — целесообразность в построении *гибридного алгоритма*, которому дадим название ***полюсно-бесполюсный метод Ньютона***. Его организация может быть, например, следующей.

Шаг 0. Задать начальное приближение x_0, способ выбора полюсов $P_k(c_k; d_k)$, критерий окончания. Положить $k := 0$.

Шаг 1. Вычислить $\tilde{x}_{k+1} = x_k - \dfrac{f(x_k)}{f'(x_k)}$,

$$\bar{x}_{k+1} = x_k - \frac{f(x_k)}{f'(x_k) - \dfrac{d_k}{c_k - x_k}}.$$

Шаг 2. Сравнить знаки $f(\tilde{x}_{k+1})$ и $f(\bar{x}_{k+1})$; если $f(\tilde{x}_{k+1})\,f(\bar{x}_{k+1}) \le 0$, то положить $x_{k+1} := \frac{1}{2}(\tilde{x}_{k+1} + \bar{x}_{k+1})$ и перейти к шагу 5.

Шаг 3. Сравнить значения $f(\tilde{x}_{k+1})$ и $f(\bar{x}_{k+1})$; если $|f(\tilde{x}_{k+1})| \le |f(\bar{x}_{k+1})|$, то положить $x_{k+1} := \tilde{x}_{k+1}$ и перейти к шагу 5.

Шаг 4. Положить $x_{k+1} := \bar{x}_{k+1}$.

Шаг 5. Сделать проверку на точность; если критерий окончания не выполняется, положить $k := k+1$ и вернуться к шагу 1; иначе остановить работу алгоритма, считая $\xi :\approx x_{k+1}$.

Теперь обратимся к полюсному методу секущих (5.37). Сначала рассмотрим пример, демонстрирующий поведение этого метода.

Пример 5.6. Для нахождения нуля той же функции $f(x) = \ln x$, что и в предыдущем примере 5.5, применим полюсный метод секущих (5.37), зафиксировав в нем абсциссу полюса $c = -2$ и, изменяя ординату d на каждой итерации k по правилу $d := d_k := f(x_{k-1})$, т.е. проводя вычисления по формулам

$$d_k = f(x_{k-1}), \quad x_{k+1} = x_k - \frac{f(x_k)}{\dfrac{d_{k-1} - f(x_k)}{x_{k-1} - x_k} + \dfrac{d_{k-1}}{2 + x_k}} \tag{5.56}$$

при $k = 1, 2, \ldots$. Беря начальное приближение $x_0 = \frac{1}{e}$ (то же, что и в примере 5.5) и начальный сдвиг $h = 10^{-4}$ (иначе, требующуюся для двухшагового метода вторую начальную точку $x_1 = x_0 + 10^{-4}$), получим результаты, показанные в табл. 5.2. Здесь же отражен и другой случай, когда берутся начальные приближения $x_0 = e$, $x_1 = x_0 - 10^{-4}$, приводящие к расходимости классического метода секущих (5.32).

Как видим, полюсный метод секущих оказывается более выигрышным по отношению к базовому для него классическому методу секущих и сравним по требуемому количеству значений функций (горнеров) с продемонстрированным выше полюсным методом Ньютона.

Таблица 5.2

Сравнительное поведение классического (5.32) *и полюсного (в модификации* (5.56)) *методов секущих*

k	Классический метод секущих		Полюсный метод секущих вида (5.56)			
	Случай $x_0 = \frac{1}{e}$, $x_1 = x_0 + 10^{-4}$				Случай $x_0 = e$, $x_1 = x_0 - 10^{-4}$	
	x_k	$f(x_k)$	x_k	$f(x_k)$	x_k	$f(x_k)$
0	0.367879441	–1.00000000	0.367879441	–1.000000000	2.718281828	1.000000000
1	0.367979441	–0.999728209	0.367979441	–0.999728209	2.718181828	0.999963211
2	0.735808880	–0.306784868	0.803474828	–0.218809422	0.993608561	–0.006411952
3	0.898656979	–0.106853876	0.955788591	–0.045218529	1.000596432	0.000596255
4	0.985691762	–0.014411588	0.998220796	–0.001780789	1.000000638	0.000000638
5	0.999260330	–0.000739944	0.999986373	–0.000013627	1.000000000	–0.000000000
6	0.999994695	–0.000005305	0.999999996	–0.000000004		
7	0.999999998	–0.000000002				

Определимся с тем, как нужно задавать координаты полюса $P(c; d)$ в полюсном методе секущих (5.37), чтобы можно было рассчитывать на его сходимость, причем более быструю, чем у основного метода секущих (5.32).

Будем исходить из геометрических соображений в предположении, что, во-первых, вторая начальная точка, т.е. x_1, лежит между точкой x_0 и корнем ξ уравнения $f(x) = 0$, а во-вторых, функция $f(x)$ имеет определенное поведение в смысле направления возрастания и выпуклости.

Рассмотрим случай, изображенный на рис. 5.15, когда $x_0 < x_1 < \xi$, $f'(x) > 0$, $f''(x) < 0$.

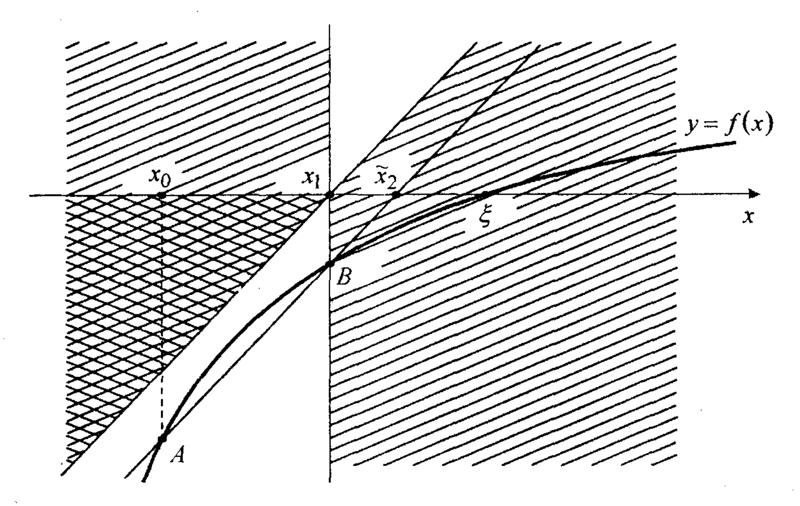


Рис. 5.15. Зоны, пригодные для выбора полюса $P(c;d)$ в полюсном методе секущих (5.37)

Проведем через точку $(x_1;0)$ прямую

$$y=\frac{f(x_0)-f(x_1)}{x_0-x_1}(x-x_1),$$

параллельную соответствующей приближениям x_0, x_1 секущей AB. Очевидно, что эта прямая и прямая $x=x_1$ разбивают всю координатную плоскость на две части такие, что в одной из них (не заштрихована) полюс P брать заведомо бессмысленно, так как формула (5.37) будет приводить к точке x_2, отстоящей от ξ дальше, чем x_1. Из двух заштрихованных центральносимметричных относительно точки $(x_1;0)$ допустимых для выбора P областей ограничимся лишь одной, расположенной левее прямой $x=x_1$. Отбросим теперь область, состоящую из точек $(x;y)$ таких, в которых одновременно $x<x_1$, $y>0$, поскольку выбор полюса P в этой области не даст выигрыша методу (5.37) по сравнению с классическим методом секущих (5.32) (на рис. 5.15 шагу по формуле секущих отвечает точка $\tilde{x}_2$).

Таким образом, приходим к тому, что в рассматриваемом случае полюс $P(c;d)$ в полюсном методе секущих целесообразно выбирать в области, выделенной на рис. 5.15. двойной штриховкой, т.е. параметры c и d должны удовлетворять неравенствам

$$c<x_1,\quad \frac{f(x_0)-f(x_1)}{x_0-x_1}(c-x_1)<d<0. \qquad (5.57)$$

При этом можно c жестко зафиксировать, а d изменять в такт с получаемыми значениями функции (как это сделано в примере 5.6, где поведение функции и начальных приближений соответствует изученному случаю), а можно изменять и ту и другую координаты полюса, например, подобно тому, как это делалось в полюсном методе Ньютона.

Аналогично рассматриваются и другие случаи взаимного поведения функции $f(x)$ и начальных приближений x_0, x_1 (читателю предлагается проделать это самостоятельно и выписать для каждого случая неравенства типа (5.57)).

УПРАЖНЕНИЯ

5.1. Докажите существование и единственность корня уравнения

$$(x-1)^2e^x-7=0.$$

5.2. Найдите промежуток локализации (единичной длины) отрицательного корня уравнения $x^3-x^2+4=0$. За сколько шагов метода половинного деления можно уточнить корень до 0.1? до 0.01? до 10^{-6}?

5.3. С точностью до 0.01 решите уравнение $\sqrt{|x-4|}-x+1=0$:
а) методом половинного деления; б) методом хорд.

5.4. С точностью до 0.001 найдите положительный корень уравнения

$$x^4-2x-4=0:$$

а) методом Ньютона; б) методом секущих.
Уточните корень, применив к результату б) два шага метода Стеффенсена.

5.5. Подготовьте алгоритм вычисления значения функции $y=\sqrt[7]{x}$ в точке $x=100$ с точностью $\varepsilon=10^{-6}$, пользуясь правилом Ньютона (§ 5.5). Сделайте два приближения.

5.6. На основе метода Ньютона запишите итерационный процесс, который позволял бы вычислять значения $\frac{1}{\sqrt{a}}$ при заданных вещественных значениях a, не производя делений.

5.7. Составьте гибридный алгоритм «секущих-половинного деления». Изучите его поведение на уравнении $x^4+(x-2)^2-8=0$, строя приближения к положительному корню из разных пар начальных точек x_0, x_1.

5.8. По аналогии с примерами 5.5, 5.6 (§ 5.7) проанализируйте сравнительное поведение приближений к корню уравнения $f(x)=0$, начинаемых со значения $x_0 = 2$ и продолжаемых методами Ньютона, секущих и полюсными методами Ньютона и секущих с подвижными полюсами, для функций: а) $f(x)=e^x - 1$, б) $f(x)=\text{arctg } x$.

5.9. Примените гибридный алгоритм «полюсно-безполюсный метод Ньютона» для нахождения корня уравнения

$$x \ln x - x - 6 = 0$$

с точностью $\varepsilon = 10^{-6}$.

ГЛАВА 6 СКАЛЯРНАЯ ЗАДАЧА О НЕПОДВИЖНОЙ ТОЧКЕ. АЛГЕБРАИЧЕСКИЕ УРАВНЕНИЯ

Задача отыскания корня нелинейного скалярного уравнения, изучавшаяся в предыдущей главе, преобразуется к эквивалентной задаче о нахождении неподвижной точки некоторого нелинейного отображения в одномерном пространстве. К последней применяется метод простых итераций (МПИ), изучаются условия и характер его сходимости. Далее рассматриваются два способа ускорения сходимости МПИ, а именно, хорошо известный Δ^2-процесс Эйткена и менее известный, но, как убедительно доказывают приводимые здесь примеры, более эффективный метод Вегстейна. На логистическом уравнении (с параметром) исследуется возможное поведение генерируемых МПИ последовательностей в случае нарушения одного из достаточных условий сходимости, дается представление о бифуркациях решений и циклов. Обсуждается специфика алгебраических уравнений и обозначаются некоторые методы их решения (в частности, рассматривается метод Бернулли, идейно близкий изучавшемуся в гл. 4 степенному методу решения частичной алгебраической проблемы собственных значений).

6.1. ЗАДАЧА О НЕПОДВИЖНОЙ ТОЧКЕ. МЕТОД ПРОСТЫХ ИТЕРАЦИЙ

Нельзя не заметить, что все расчетные формулы, определяющие уже изученные методы решения скалярных уравнений вида $f(x)=0$, т.е. (5.1), такие как метод Ньютона (5.14) и его модификации (5.25), (5.27), (5.29), (5.31), (5.36), имеют вид

$$x_{k+1} = \varphi(x_k), \quad k = 0, 1, 2, \dots , \tag{6.1}$$

где $\varphi(x)$ — некоторая функция, для каждого метода своя, так или иначе связанная с исходной функцией $f(x)$. Попытаемся понять, каким требованиям должна удовлетворять функция $\varphi(x)$, чтобы последовательность (x_k), определяемая этим самым общим одношаговым итерационным способом (6.1), называемым

методом простых итераций*), была сходящейся, и как построить функцию $\varphi(x)$ по функции $f(x)$, чтобы эта последовательность сходилась к корню данного уравнения $f(x)=0$.

Сразу отметим, что функцию $\varphi(x)$ будем считать непрерывной в исследуемой области оси Ox. Поэтому, если определяемая формулой (6.1) последовательность (x_k) окажется сходящейся к некоторому числу ξ, то, переходя к пределу в равенстве (6.1), получаем

$$\xi=\varphi(\xi), \tag{6.2}$$

т.е. $\xi=\lim\limits_{k\to\infty} x_k$ — корень уравнения

$$x=\varphi(x). \tag{6.3}$$

Решение уравнений именно вида (6.3), представляет самостоятельный интерес; нахождение их корней называется ***задачей о неподвижной точке***. Это название связано с тем, что точка ξ при отображении с помощью φ из $\mathbf{R}_1$ в $\mathbf{R}_1$ остается на месте (разумеется, если таковая существует).

Существование и единственность корня уравнения (6.3) основывается на ***принципе сжимающих отображений*** или, иначе, ***принципе неподвижной точки***.

Определение 6.1. *Непрерывная функция $\varphi(x)$ называется* ***сжимающей*** (***функцией сжатия***) *на отрезке $[a,b]$, если:*

1) $\varphi(x)\in[a,b] \quad \forall x\in[a,b]$;

2) $\exists q\in(0,1)\colon |\varphi(x_2)-\varphi(x_1)|\le q|x_2-x_1| \quad \forall x_1,x_2\in[a,b]$.

Графическое толкование применения сжимающего отображения φ (как функции множества) к промежутку сжатия $[a,b]$ предоставляет рис. 6.1.

---

*) Другие названия: ***метод итераций***, ***метод последовательных приближений***. Далее, как и в гл. 3, будем также использовать аббревиатуру МПИ.

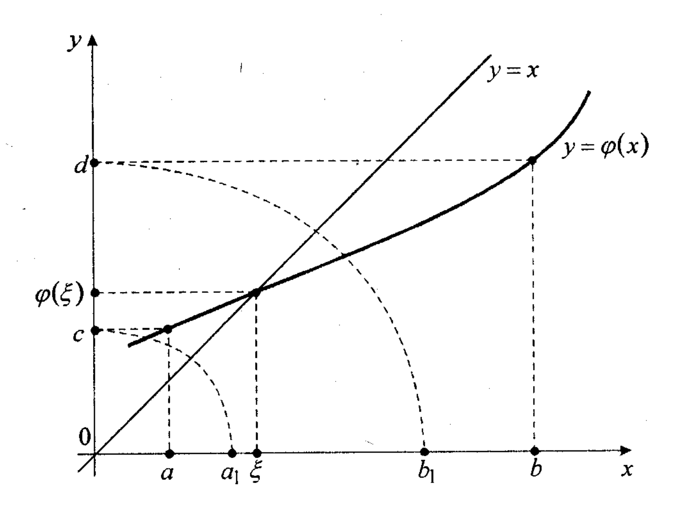


Рис. 6.1. Сжатие отрезка $[a,b]$ возрастающей функцией $\varphi(x)$

Как видно из этого рисунка, если $[a,b]$ рассматривать как область определения функции сжатия $\varphi(x)$, то соответствующая ей область значений $[\varphi(a),\varphi(b)]$ на оси ординат (отрезок $[c,d]$), будучи перенесенным на ось абсцисс (отрезок $[a_1,b_1]$), целиком содержится в $[a,b]$. Применяя к $[a_1,b_1]$ те же рассуждения, что и к $[a,b]$, получим $[a_2,b_2]\subset[a_1,b_1]$, и т.д. В итоге образуется бесконечная последовательность вложенных отрезков

$$[a,b]\supset[a_1,b_1]\supset[a_2,b_2]\supset\ldots\supset[a_k,b_k]\supset\ldots,$$

причем их длины убывают по закону

$$b_k-a_k\le q^k(b-a)\underset{k\to\infty}{\longrightarrow}0.$$

Следовательно, в условиях сжатия эта последовательность имеет единственную общую точку (ξ), которая переходит сама в себя, т.е. является неподвижной точкой отображения φ. При этом, очевидно, последовательность (a_k) левых концов этих промежутков монотонно сходится к ξ слева, а последовательность (b_k) правых концов — справа. Так как при условии возрастания $\varphi(x)$, как это показано на рис.6.1, в этом процессе

$$a_1=\varphi(a),\quad a_2=\varphi(a_1),\quad a_3=\varphi(a_2),\quad \ldots,$$
$$b_1=\varphi(b),\quad b_2=\varphi(b_1),\quad b_3=\varphi(b_2),\quad \ldots,$$

то можно утверждать, что МПИ (6.1) будет давать монотонно сходящуюся к ξ последовательность, если ее начинать с $x_0 = a$ или с $x_0 = b$. Так же монотонно возрастающая и монотонно убывающая последовательности приближений будут получаться по формуле (6.1) и в случаях, когда за x_0 будет браться любая точка из промежутков $[a, \xi)$ и $(\xi, b]$ соответственно.

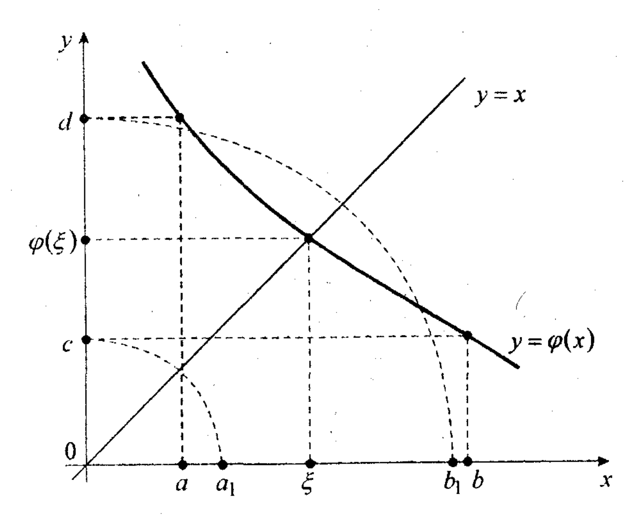


Рис. 6.2. Сжатие отрезка $[a, b]$ убывающей функцией $\varphi(x)$

При условии убывания сжимающей функции $\varphi(x)$, т.е. в случае, изображенном на рис. 6.2, начинающиеся с концов a и b промежутка сжатия последовательности выстраиваются следующим образом:

$$a,\quad b_1 = \varphi(a),\quad a_2 = \varphi(b_1),\quad b_3 = \varphi(a_2),\quad \ldots,$$

$$b,\quad a_1 = \varphi(b),\quad b_2 = \varphi(a_1),\quad a_3 = \varphi(b_2),\quad \ldots.$$

Каждая из них сходится к неподвижной точке ξ, и элементы каждой из этих последовательностей с удалением от начала дают все более хорошие приближения то с недостатком, то с избытком. Такую сходимость к ξ имеет и любая другая последовательность (x_k), получаемая по формуле (6.1) при любом $x_0 \in [a, b]$. Отсюда другой термин, применяемый к неподвижной точке ξ, — ***центр итерации*** [140].

Более удобно иллюстрировать геометрически поведение итерационной последовательности (x_k), определяемой МПИ

(6.1), не отмечая значения $\varphi(x_k)$ на оси ординат, а отражая их на ось абсцисс с помощью биссектрисы координатного угла $y = x$. Такие иллюстрации для случаев монотонного возрастания (ломаная типа «ступеньки») и монотонного убывания (ломаная типа «спираль») сжимающей функции $\varphi(x)$ показаны на рис. 6.3 и 6.4 соответственно (обоснуйте!).

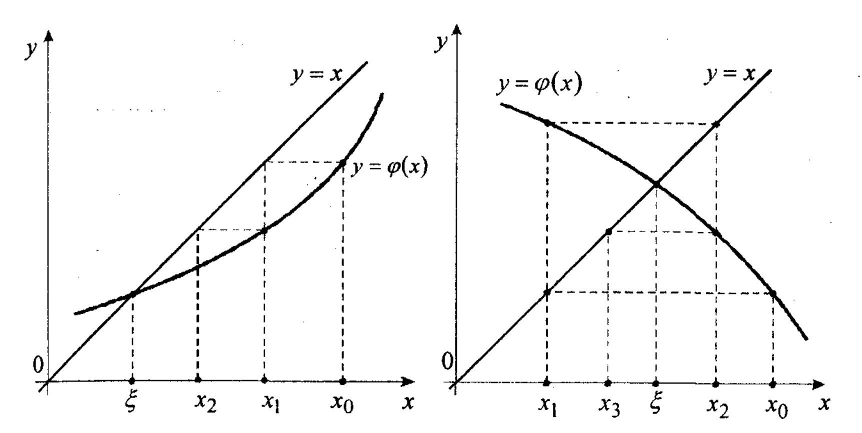


Рис. 6.3. Монотонные приближения к корню

Рис. 6.4. Двусторонние приближения к корню

Итогом проведенных выше рассуждений является следующий **вывод:**

*если на некотором промежутке $[a, b]$ функция $\varphi(x)$ удовлетворяет условиям сжатия, зафиксированным определением 6.1, то: 1) уравнение (6.3) имеет и притом единственный корень $\xi \in [a, b]$; 2) к этому корню со скоростью геометрической прогрессии сходится определяемая МПИ (6.1) последовательность (x_k), начинающаяся с любого $x_0 \in [a, b]$, причем скорость сходимости тем выше, чем меньше **коэффициент сжатия***) $q \in (0, 1)$; 3) если функция $\varphi(x)$ монотонно возрастает на $[a, b]$, то приближения x_k к ξ также будут монотонными, если же $\varphi(x)$ убывает, то процесс (6.1) порождает двусторонние приближения к корню ξ.*

---

*) Иначе, ***константа Липшица***, в соответствии с применяемым к неравенству $|\varphi(x_2) - \varphi(x_1)| \le q|x_2 - x_1|$ термином ***условие Липшица*** (или ***Коши–Липшица***, если заведомо $q \in (0, 1)$).

Этот вывод имеет, в основном, качественный характер. Дальнейшие исследования будут направлены на выявление конструктивных требований к $\varphi(x)$, обеспечивающих для нее выполнение условий сжатия, и получение оценок близости генерируемых МПИ (6.1) приближений x_k к неподвижной точке ξ.

Теорема 6.1. *Пусть функция $\varphi(x)$ определена и дифференцируема на отрезке $[a, b]$. Тогда, если выполняются условия:*

1) $\varphi(x) \in [a, b] \quad \forall\, x \in [a, b]$,

2) $\exists\, q : |\varphi'(x)| \le q < 1 \quad \forall x \in (a, b)$,

то уравнение (6.3) *имеет и притом единственный на* $[a, b]$ *корень* ξ; *к этому корню сходится определяемая методом простых итераций* (6.1) *последовательность* (x_k), *начинающаяся с любого* $x_0 \in [a, b]$; *при этом справедливы оценки погрешности* $(\forall k \in N)$:

$$|\xi - x_k| \le \frac{q}{1-q}|x_k - x_{k-1}|, \tag{6.4}$$

$$|\xi - x_k| \le \frac{q^k}{1-q}|x_1 - x_0|. \tag{6.5}$$

Доказательство. Взяв произвольное x_0 из $[a, b]$, согласно первому требованию к $\varphi(x)$, заключаем, что значение $x_1 = \varphi(x_0)$ также принадлежит $[a, b]$. По индукции можно утверждать, что все члены последовательности (x_k), генерируемой МПИ (6.1), благодаря условию 1) не могут покинуть отрезок $[a, b]$.

Вычтем из равенства (6.1) такое же равенство

$$x_k = \varphi(x_{k-1})$$

и к правой части полученного равенства

$$x_{k+1} - x_k = \varphi(x_k) - \varphi(x_{k-1})$$

применим формулу Лагранжа, согласно которой на интервале, определяемом точками x_{k-1} и x_k (а значит, на интервале (a, b)), найдется точка θ_k такая, что

$$\varphi(x_k) - \varphi(x_{k-1}) = \varphi'(\theta_k)(x_k - x_{k-1}).$$

Следовательно,

$$x_{k+1} - x_k = \varphi'(\theta_k)(x_k - x_{k-1})$$

и, в силу условия 2), справедливо неравенство

$$|x_{k+1} - x_k| \le q|x_k - x_{k-1}|, \tag{6.6}$$

которое можно расценивать как выполнение главного условия сжатия на элементах итерационной последовательности (x_k) с коэффициентом сжатия q. С другой стороны, неравенство (6.6) указывает на факт и скорость сближения членов этой последовательности.

Для разностей соседних более удаленных от начала членов последовательности (x_k) на основе (6.6) получаем:

$$\begin{aligned} &|x_{k+2} - x_{k+1}| \le q|x_{k+1} - x_k| \le q^2|x_k - x_{k-1}|; \\ &|x_{k+3} - x_{k+2}| \le q|x_{k+2} - x_{k+1}| \le q^3|x_k - x_{k-1}|; \\ &\dots\dots\dots\dots\dots\dots \\ &|x_{k+i} - x_{k+i-1}| \le q^i|x_k - x_{k-1}| \quad \forall i \in \mathbf{N}_0, \quad k \in \mathbf{N}. \end{aligned} \tag{6.7}$$

Используя эти неравенства, оценим близость между x_{k+m} (где $m \in \mathbf{N}$) и x_k, вычитая и прибавляя все промежуточные члены x_{k+m-1}, $x_{k+m-1}, x_{k+m-2}, \dots, x_{k+1}$ и применяя свойство «модуль суммы не превосходит суммы модулей». Имеем:

$$|x_{k+m} - x_k| \le |x_{k+m} - x_{k+m-1}| + |x_{k+m-1} - x_{k+m-2}| + \dots$$

$$\dots + |x_{k+2} - x_{k+1}| + |x_{k+1} - x_k| \le \left(q^m + q^{m-1} + \dots + q^2 + q\right)|x_k - x_{k-1}| =$$

$$= \frac{q - q^{m+1}}{1-q}|x_k - x_{k-1}|.$$

Но, в свою очередь,

$$|x_k - x_{k-1}| \le q^{k-1}|x_1 - x_0|, \tag{6.8}$$

что можно установить либо итерируя неравенство (6.6) при $k = 1, 2, \dots$, либо непосредственно из (6.7), полагая $k = 1$, а затем $i = k - 1$. Подставляя (6.8) в полученную выше оценку

$$|x_{k+m} - x_k| \le \frac{q}{1-q}\left(1 - q^m\right)|x_k - x_{k-1}|, \tag{6.9}$$

имеем

$$|x_{k+m} - x_k| \le \frac{q^k}{1-q}\left(1 - q^m\right)|x_1 - x_0|. \qquad (6.10)$$

Поскольку правая часть неравенства (6.10) при фиксированном $m \in \mathbf{N}$ и $k \to \infty$ стремится к нулю, (x_k) — последовательность Коши и, в силу замкнутости отрезка, имеет предел $\xi \in [a, b]$. Так как дифференцируемая функция непрерывна, то этот предел — корень уравнения (6.3). Его единственность на $[a, b]$ доказывается от противного: предположив, что наряду с ξ есть другой корень $\tau \in [a, b]$, т.е. имеет место равенство $\tau = \varphi(\tau)$, а значит, $\xi - \tau = \varphi(\xi) - \varphi(\tau)$, по формуле Лагранжа получаем

$$\exists \theta \in (a, b): \;\; \xi - \tau = \varphi'(\theta)(\xi - \tau);$$

последнее же равенство возможно лишь при $\xi = \tau$, поскольку по условию производная не может быть равна единице.

Перейдя к пределу в неравенствах (6.9) и (6.10) при $m \to \infty$, получаем оценки (6.4) и (6.5) соответственно, что и завершает доказательство теоремы.

Анализируя условия и доказательство теоремы 6.1, нельзя не заметить, что основным требованием к функции $\varphi(x)$, обеспечивающим сходимость МПИ (6.1) с оценками (6.4), (6.5), является условие малости модуля производной. Требование же $\varphi(x) \in [a, b]$ нужно лишь постольку, поскольку оно должно обеспечить попадание значений $\varphi(x)$ на промежуток, где выполняются другие требования. Очевидно, что достаточно его выполнения только на элементах итерационной последовательности (x_k), и практически это может проверяться непосредственно в процессе счета по формуле (6.1). Теоретически же, когда $[a, b]$ — не вся числовая ось (тогда надобности в этом условии нет), имеется несколько возможностей заменить требование $\varphi(x) \in [a, b]$ более конструктивным, т.е. априори проверяемым условием.

Будем исходить из того, что имеется некоторая точка x_0, которую можно взять в качестве начального приближения и в окрестности которой функция $\varphi(x)$ дифференцируема и имеет малую по модулю производную. Тогда существование и единственность корня ξ уравнения (6.3) и сходимость к нему начатого с x_0 процесса (6.1) можно связать с величиной r радиуса этой окрестности точки x_0.

Теорема 6.2. *Пусть на отрезке $[x_0 - r, \, x_0 + r]$ функция $\varphi(x)$ определена, дифференцируема и ее производная удовлетворяет неравенству*

$$|\varphi'(x)| \le q < 1. \qquad (6.11)$$

Тогда, если величина r такова, что

$$|x_0 - \varphi(x_0)| \le r(1-q), \qquad (6.12)$$

то на $[x_0 - r, \, x_0 + r]$ имеется единственный корень ξ уравнения (6.3), *и к нему сходится начатый с x_0 метод простых итераций* (6.1) *с оценками погрешности* (6.4), (6.5).

Доказательство. Так как $1 - q < 1$, то, в силу неравенства (6.12), $x_1 = \varphi(x_0) \in [x_0 - r, \, x_0 + r]$. Предположив, что $x_k = \varphi(x_{k-1})$ принадлежит r-окрестности точки x_0 при некотором $k \in \mathbf{N}$, покажем, что там же будет и $x_{k+1} = \varphi(x_k)$. Действительно, поскольку, согласно предположению, на отрезке $[x_0 - x_k, \, x_0 + x_k] \subset [x_0 - r, \, x_0 + r]$ функция $\varphi(x)$ определена и дифференцируема, на этом отрезке найдется точка μ_k такая, что

$$x_{k+1} - x_0 = \varphi(x_k) - \varphi(x_0) + \varphi(x_0) - x_0 =$$
$$= \varphi'(\mu_k)(x_k - x_0) + \varphi(x_0) - x_0.$$

Отсюда, переходя к модулям, получаем неравенство

$$|x_{k+1} - x_0| \le qr + r(1-q) = r,$$

означающее, что $x_{k+1} \in [x_0 - r, \, x_0 + r]$.

Таким образом, на элементах итерационной последовательности (x_k) выполняется первое требование теоремы 6.1 для отрезка $[a, b]$ с $a := x_0 - r$, $b := x_0 + r$, и вместе с требованием (6.11) оно обеспечивает справедливость заключения теоремы.

Полученные для МПИ простые оценки погрешности (6.4) и (6.5) можно использовать в практических вычислениях как для завершения итерационного процесса (6.1) по правилу:

$$|x_k - x_{k-1}| \le \frac{1-q}{q}\varepsilon \;\; \Rightarrow \;\; \xi := x_k (\pm \varepsilon), \qquad (6.13)$$

так и для предварительной прикидки числа итераций, достаточного для получения корня с заданной точностью ε:

$$\frac{q^k}{1-q}|x_1 - x_0| \le \varepsilon \;\; \Leftrightarrow \;\; k \ge \frac{1}{\ln q} \ln \frac{\varepsilon(1-q)}{|x_0 - \varphi(x_0)|}$$

(или $k \ge \dfrac{\ln\varepsilon - \ln r}{\ln q}$ в условиях теоремы 6.2).

Замечание 6.1. Легко видеть, что часто применяемый на практике простой критерий окончания процесса итераций (6.1) по выполнении неравенства $|x_k - x_{k-1}| \le \varepsilon$ обоснован в двух случаях: когда $-1 < \varphi'(x_k) \le 0$, т.е. МПИ дает двусторонние приближения (корень ξ всегда «зажат» между любыми двумя соседними приближениями, расстояние между которыми служит эффективной оценкой погрешности, см. рис. 6.4), и когда $0 \le \varphi'(x_k) \le q < 1$, но при этом $q \le \frac{1}{2}$ (тогда $\frac{1-q}{q} \ge 1$, и фигурирующее в (6.13) неравенство будет заведомо выполнено). Использование этого упрощенного критерия окончания при значениях q, близких к концам интервала $(0,1)$ чревато либо лишними итерациями, либо недоитерированием.

Обратимся к связи между уравнением $f(x) = 0$ (5.1) и задачей о неподвижной точке — уравнением (6.3). Переписав (6.3) в виде

$$x - \varphi(x) = 0, \tag{6.14}$$

можно сказать, что это есть уравнение (5.1) с $f(x) := x - \varphi(x)$, и применять к (6.14) все рассмотренные в предыдущих параграфах рассуждения и методы. Приведение уравнения (5.1) к виду (6.3) можно осуществлять множеством способов, но при этом всегда следует помнить, что это приведение нужно выполнять так, чтобы полученное уравнение соответствующего вида было не только эквивалентным (5.1), но и пригодным для проведения итераций, т.е. удовлетворяло оговоренным в теоремах 6.1, 6.2 условиям. Кроме того, попутно могут учитываться такие требования к получающемуся при этом методу итераций, как простота расчетной формулы, быстрота сходимости (малость q), характер сходимости (монотонность или двусторонность приближений). Если уравнение (5.1) имеет несколько корней, то для нахождения каждого из них формируется своя задача о неподвижной точке.

Иногда преобразование уравнения (5.1) к виду (6.3) не вызывает больших затруднений и выполняется непосредственно.

Например, уравнение

$$4 - 2x - \sin x = 0$$

достаточно записать в виде

$$x = 2 - 0.5\sin x,$$

чтобы сказать, что оно имеет и притом единственное в $\mathbf{R}$ решение, к которому сходится при любом $x_0 \in \mathbf{R}$ последовательность

$$x_{k+1} = 2 - 0.5\sin x_k$$

со скоростью геометрической прогрессии со знаменателем $q \le 0.5$, поскольку функция $\varphi(x) = 2 - 0.5\sin x$ имеет производную $\varphi'(x) = -0.5\cos x$, абсолютная величина которой не превосходит 0.5 при любых $x \in \mathbf{R}$.

В общем случае переход от (5.1) к (6.3) осуществляют так: умножают левую и правую части уравнения (5.1) на отличный от нуля параметр $-\lambda$ и к обеим частям прибавляют по x; в результате получается равносильное (5.1) уравнение

$$x = x - \lambda f(x), \tag{6.15}$$

которое имеет вид (6.3) с $\varphi(x) := x - \lambda f(x)$. Далее параметр λ подбирается таким, чтобы производная $\varphi'(x) = 1 - \lambda f'(x)$ в нужной области была малой по модулю (а если надо, то чтобы еще имела определенный знак).

Конкретные рекомендации по фиксированию λ в (6.15) могут быть даны в случае, когда, например, известны оценки сверху и снизу для производной исходной функции $f(x)$. А именно, пусть

$$0 < \alpha \le f'(x) \le \gamma < \infty$$

(если производная $f'(x)$ отрицательна, можно заменить уравнение $f(x) = 0$ на уравнение $-f(x) = 0$, т.е. работать с функцией $-f(x)$). Тогда, соответственно,

$$1 - \lambda\gamma \le \varphi'(x) \le 1 - \lambda\alpha, \tag{6.16}$$

и значит,

$$|\varphi'(x)| \le q(\lambda) := \max\{|1 - \lambda\alpha|, |1 - \lambda\gamma|\}.$$

Анализируя двойное неравенство (6.16), можно увидеть, что при любых $\lambda \in \left(0, \frac{2}{\gamma}\right)$ будет $q(\lambda) < 1$. В частности, при $\lambda = \frac{1}{\gamma}$ имеет место неравенство

$$0 \le \varphi'(x) \le 1 - \frac{\alpha}{\gamma} < 1,$$

обеспечивающее монотонную сходимость соответствующего МПИ со скоростью, определяемой оценками (6.4), (6.5) при $q = 1 - \frac{\alpha}{\gamma}$. Оптимальным же значением λ является $\lambda = \lambda_0 := \frac{2}{\alpha + \gamma}$. При этом значении λ границы неравенства (6.16) таковы:

$$1 - \lambda\gamma = \frac{\alpha - \gamma}{\alpha + \gamma}, \quad 1 - \lambda\alpha = \frac{\gamma - \alpha}{\alpha + \gamma},$$

т.е. максимум $|\varphi'(x)|$, равный $q(\lambda_0)=\dfrac{\gamma-\alpha}{\alpha+\gamma}$, достигается на каждом из элементов двухэлементного множества $\{|1-\lambda\alpha|, |1-\lambda\gamma|\}$.

В отсутствие нужных оценок для $f'(x)$ можно предложить следующие рассуждения. Если известно, что искомый корень находится в окрестности заданной точки x_0, где производная меняется не очень быстро, возьмем λ таким, чтобы $\varphi'(x_0)=0$. Тогда по непрерывности производная должна остаться малой в окрестности x_0, т.е. можно рассчитывать на сходимость получающегося при этом итерационного процесса. Имеем:

$$1-\lambda f'(x_0)=0 \;\Rightarrow\; \lambda=\frac{1}{f'(x_0)}.$$

Подставляя это λ в (6.15), получаем уравнение вида (6.3)

$$x=x-\frac{f(x)}{f'(x_0)},$$

и соответствующий ему МПИ (6.1) определяется формулой

$$x_{k+1}=x_k-\frac{f(x_k)}{f'(x_0)},$$

в которой узнаём ***модифицированный метод Ньютона****) (5.27).

6.2. УСКОРЕНИЕ СХОДИМОСТИ ПОСЛЕДОВАТЕЛЬНЫХ ПРИБЛИЖЕНИЙ

Как явствует из предыдущего параграфа, МПИ (6.1) имеет лишь линейную сходимость, причем в случаях, когда производная функции $\varphi(x)$ близка к единице, эта сходимость может быть весьма медленной.

Один путь ускорения сходимости МПИ — построение

*) В [138, 139] метод простых итераций, примененный к уравнению (6.15), т.е. процесс вида

$$x_{k+1}=x_k-\alpha f(x_k),$$

при условии, что $\operatorname{sign}\alpha=\operatorname{sign} f'(x_k)$, называют ***методом хорд*** (или параллельных хорд), и модифицированный метод Ньютона считают частным случаем этого метода.

нестационарных процессов на основе МПИ. На стадии приведения уравнения (5.1) к задаче о неподвижной точке (6.3) можно подменить (5.1) не однопараметрическим семейством уравнений (6.15), а последовательностью уравнений

$$x=x-\lambda_k f(x),\quad k=0,1,2,\dots,$$

и в соответствующей таким уравнениям формуле МПИ

$$x_{k+1}=x_k-\lambda_k f(x_k) \tag{6.17}$$

параметры λ_k подбирать так, чтобы при этом учитывалась информация, получаемая на предыдущем шаге. Если, например, выбор параметра λ_k подчинить соображениям, что $\varphi'(x_k):=1-\lambda_k f'(x_k)=0$, т.е. взять $\lambda_k=\dfrac{1}{f'(x_k)}$, то (6.17) будет определять основной метод Ньютона, сходящийся, как уже известно (см. § 5.4), квадратично. Можно выбирать и другие стратегии фиксирования λ_k, но вряд ли здесь следует ожидать нечто принципиально новое, что нельзя более естественно получить с помощью формулы Тейлора и интерполяционных формул.*)

Другой путь — это алгоритмическое построение последовательностей, так или иначе «паразитирующих» на последовательности приближений МПИ (6.1), т.е. получаемых с помощью несложных арифметических манипуляций над несколькими членами последовательности (x_k) и в результате имеющих более быструю сходимость. Для всех таких методов характерны *многошаговость*, *экономичность* (поскольку более быстрая сходимость по сравнению с базовой последовательностью достигается без дополнительного вычисления значений функций), а также *сложность исследования* условий и скорости сходимости, отсюда — *отсутствие эффективных априорных оценок погрешностей*. Возможны ситуации, когда новый метод окажется сходящимся, в то время как базовый для него МПИ расходится.

Рассмотрим два таких метода ускорения сходимости последовательности (6.1), не делая попыток их строгого обоснования, а ограничиваясь рациональными рассуждениями при их выводе, а также наглядными примерами, демонстрирующими их эффективность. *Наличие неподвижной точки ξ и дифференцируемость функции $\varphi(x)$ далее всюду предполагается.*

*) Довольно обширная и глубокая теория итерационных методов, в основу которой положено понятие ***итерационной функции***, содержится в монографии известного американского математика Дж. Трауба [176].

6.2.1. Δ^2-ПРОЦЕСС ЭЙТКЕНА*)

Пусть (x_k) — последовательность, получаемая по форму (6.1). Вычитая (6.1) из (6.2), имеем

$$\xi - x_{k+1} = \varphi(\xi) - \varphi(x_k),$$

а уменьшив здесь индекс на единицу, получаем

$$\xi - x_k = \varphi(\xi) - \varphi(x_{k-1}).$$

К правым частям этих равенств применим формулу Лагранжа, согласно которой найдутся точки c_k и c_{k-1} такие, что

$$\varphi(\xi) - \varphi(x_k) = \varphi'(c_k)(\xi - x_k)$$

и

$$\varphi(\xi) - \varphi(x_{k-1}) = \varphi'(c_{k-1})(\xi - x_{k-1}).$$

Таким образом, имеют место следующие связи между ошибками соседних приближений:

$$\xi - x_{k+1} = \varphi'(c_k)(\xi - x_k), \quad \xi - x_k = \varphi'(c_{k-1})(\xi - x_{k-1}).$$

Предположим, что в той окрестности корня ξ, в которой находятся точки x_{k-1} и x_k, производная $\varphi'(x)$ меняется не очень быстро. Это допущение позволяет считать, что

$$\varphi'(c_k) \approx \varphi'(c_{k-1}) \approx \eta$$

(где η — некоторое число), и значит,

$$\xi - x_{k+1} \approx \eta(\xi - x_k), \quad \xi - x_k \approx \eta(\xi - x_{k-1}).$$

Беря отношение этих приближенных равенств, избавляемся от η:

$$\frac{\xi - x_{k+1}}{\xi - x_k} \approx \frac{\xi - x_k}{\xi - x_{k-1}}, \tag{6.18}$$

и разрешаем полученное приближенное уравнение относительно неизвестной величины ξ:

$$\xi \approx \frac{x_{k+1}x_{k-1} - x_k^2}{x_{k+1} - 2x_k + x_{k-1}}. \tag{6.19}$$

---

*) Метод (читается: «дельта-два-процесс») заложен в публикации A.Aitken 1931 г. Первоначально был предназначен для улучшения метода Бернулли решения алгебраических уравнений (см. далее § 6.4). Развитию той же идеи посвятил свою статью в 1933 г. I.F.Steffensen, в связи с чем, например, в книге [192], где изучается сходимость метода и дается его обобщение, он называется ***итерационный процесс Эйткена–Стеффенсена***.

Приближенное выражение корня ξ по формуле (6.19) можно использовать на завершающем этапе применения метода простых итераций (6.1), чтобы получить более точное значение ξ с помощью трех последних членов последовательности (x_k). В развитие же метода обозначим правую часть приближенного равенства (6.19) через $\widetilde{x}_{k+1}$ и придадим его выражению другой вид:

$$\widetilde{x}_{k+1} := \frac{x_{k+1}x_{k-1} - x_k^2}{x_{k+1} - 2x_k + x_{k-1}} = x_{k+1} - \frac{(x_{k+1} - x_k)^2}{x_{k+1} - 2x_k + x_{k-1}}.$$

Более коротко это записывается так:

$$\widetilde{x}_{k+1} = x_{k+1} - \frac{(\Delta x_k)^2}{\Delta^2 x_{k-1}}, \tag{6.20}$$

где $\Delta x_k := x_{k+1} - x_k$, $\Delta^2 x_{k-1} := \Delta x_k - \Delta x_{k-1} = x_{k+1} - 2x_k + x_{k-1}$ — так называемые ***конечные разности первого и второго порядков*** соответственно (подробнее о конечных разностях см. далее в § 8.4). Отсюда название (6.20) ***Δ^2-преобразование*** или ***Δ^2-процесс Эйткена***.*)

Организация вычислений на основе этого преобразования может быть различной. Наиболее целесообразным считается применение ***Δ^2–ускорения*** (6.20) через два шага МПИ на третий**). Поскольку в этом комбинированном методе нахождения корня ξ уравнения (6.3) участвует хорошо изученный МПИ (6.1), для останова процесса вычислений можно использовать в подходящей его фазе вполне надежный критерий (6.13).

Примером реализации такого метода может служить следующий алгоритм.

Δ^2–алгоритм Эйткена

Шаг 0. Ввод x_0 (начального приближения), $\varphi(x)$ (исход-

---

*) Вместо (6.20) для Δ^2-преобразования Эйткена используют и другое представление: $\widetilde{x}_{k+1} = x_{k-1} - \dfrac{(\Delta x_{k-1})^2}{\Delta^2 x_{k-1}}$ (убедитесь в его эквивалентности (6.20)).

**) Применение Δ^2- преобразования менее, чем через два шага МПИ, необоснованно, так как оно должно применяться к трем последовательным членам л и н е й н о сходящейся последовательности.

9 Б-223

ной функции), q (оценки модуля производной), ε (допустимой абсолютной погрешности).

Шаг 1. Вычисление значений: $x_1 := \varphi(x_0)$, $x_2 := \varphi(x_1)$.

Шаг 2. Δ^2-ускорение: $\widetilde{x}_2 := \dfrac{x_0 x_2 - x_1^2}{x_2 - 2x_1 + x_0}$.

Шаг 3. Вычисление контрольного значения: $x_3 := \varphi(\widetilde{x}_2)$.

Шаг 4. Проверка на точность: если $|x_3 - \widetilde{x}_2| > \dfrac{1-q}{q}\varepsilon$, то положить $x_0 := \widetilde{x}_2$, $x_1 := x_3$, вычислить $x_2 := \varphi(x_1)$ и вернуться к шагу 2.

Шаг 5. Положить $\xi :\approx x_3$ (с точностью ε).

Шаг ускорения по методу Эйткена (6.20) на базе последовательности (x_k), получаемой МПИ (6.1), имеет простую геометрическую интерпретацию (рис. 6.5 и рис. 6.6 для случаев $-1 < \varphi'(x) < 0$ и $0 < \varphi'(x) < 1$ соответственно). Рассмотрим ее.

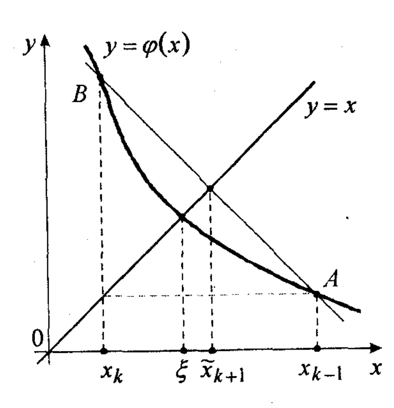


Рис. 6.5. *Ускорение по методу Эйткена (случай убывающей функции* $\varphi(x)$*)*

Рис. 6.6. *Ускорение по методу Эйткена (случай возрастающей функции* $\varphi(x)$*)*

Проведем прямую (хорду, секущую) через точки $A(x_{k-1}; \varphi(x_{k-1}))$ и $B(x_k; \varphi(x_k))$ кривой $y = \varphi(x)$. Тогда абсцисса точки пересечения этой прямой с прямой $y = x$ как раз и будет определяемая Δ^2-преобразованием Эйткена (6.20) точка $\widetilde{x}_{k+1}$ оси Ox.

Действительно, в силу (6.1), координаты точек A и B

можно записать иначе: $A(x_{k-1}; x_k)$, $B(x_k; x_{k+1})$. Значит, уравнение прямой (AB) имеет вид

$$\frac{y - x_k}{x_{k+1} - x_k} = \frac{x - x_{k-1}}{x_k - x_{k-1}}.$$

Рассматривая его совместно с уравнением $y = x$, т.е. заменяя в нем y и x на $\widetilde{x}_{k+1}$, получаем то же выражение

$$\widetilde{x}_{k+1} = \frac{x_{k+1}x_{k-1} - x_k^2}{x_{k+1} - 2x_k + x_{k-1}},$$

от которого пришли к (6.20).

Такая интерпретация Δ^2-процесса Эйткена наталкивает на мысль о его возможной связи с изученным ранее методом секущих (5.32).

Полагая $f(x) := x - \varphi(x)$, применим к этой функции формулу (5.32), считая при этом, что требуемые в (5.32) значения элементов последовательности (x_k) получаются не по той же формуле (5.32), а с помощью МПИ (6.1). Имеем:

$$x_k - \frac{f(x_k)(x_{k-1} - x_k)}{f(x_{k-1}) - f(x_k)} = x_k - \frac{(x_k - \varphi(x_k))(x_{k-1} - x_k)}{x_{k-1} - \varphi(x_{k-1}) - x_k + \varphi(x_k)} =$$

$$= x_k - \frac{(x_k - x_{k+1})(x_{k-1} - x_k)}{x_{k+1} - 2x_k + x_{k-1}} = \frac{x_{k+1}x_{k-1} - x_k^2}{x_{k+1} - 2x_k + x_{k-1}} = \widetilde{x}_{k+1}.$$

Таким образом, один шаг Δ^2-ускорения МПИ по методу Эйткена совпадает с одним шагом метода секущих, примененного к той же паре последних точек из последовательности МПИ.

Сделанное наблюдение в свете известных о методе секущих сведений позволяет с большой осторожностью судить об эффекте ускорения сходимости, который может принести использование метода Эйткена. Надежность его, очевидно, выше в случае, когда $\varphi'(x) \in (-1, 0)$, применение более актуально, если $q > \frac{1}{2}$, и ускорение тем эффективней, чем меньше $|\varphi''(x)|$ в окрестности корня ξ.

Применяя метод Эйткена, не следует забывать о проблеме своевременного прерывания счета из-за потерь точности при вычитании близких чисел. Подключение Δ^2-ускорения на ранней стадии МПИ, когда x_0 далеко от ξ, может привести к расходимости процесса, по крайней мере, в случае, когда $\varphi'(x) > 0$ (представьте эту ситуацию, рассматривая рис. 6.6). В то же вре-

мя, иногда с помощью метода Эйткена можно получить сходимость в условиях, когда МПИ (6.1) расходится (см., например, рис. 6.7, где $|\xi - x_k| > |\xi - x_{k-1}|$, но $|\xi - \tilde{x}_{k+1}| < |\xi - x_k|$).

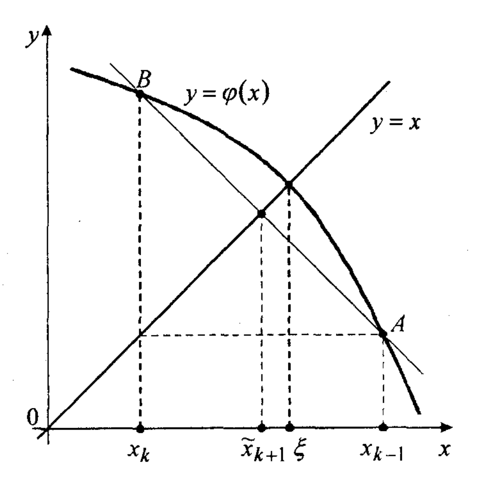


Рис. 6.7. Демонстрация возможной сходимости последовательности Эйткена $(\tilde{x}_k)$ при расходимости последовательности МПИ (x_k)

Замечание 6.2. Δ^2-преобразование Эйткена применимо не только к последовательности приближений (6.1), но и к любым другим последовательностям, сходящимся со скоростью геометрической прогрессии. Действительно, пусть (x_k) — некоторая последовательность, линейно сходящаяся к предельной точке ξ. Тогда можно считать, что разность $\xi - x_k$ изменяется по закону геометрической прогрессии, т.е. существуют такие постоянная $\nu \in (0,1)$ и слабо изменяющаяся варианта $C_k \approx C$, что

$$\xi - x_{k-1} \approx C\nu^{k-1}, \quad \xi - x_k \approx C\nu^k, \quad \xi - x_{k+1} \approx C\nu^{k+1}.$$

Отсюда получаем приближенные равенства

$$\frac{\xi - x_k}{\xi - x_{k-1}} \approx \nu, \quad \frac{\xi - x_{k+1}}{\xi - x_k} \approx \nu,$$

следствием которых является равенство (6.18), в итоге приводящее к формуле (6.20). Такой подход к выводу Δ^2-метода Эйткена позволяет использовать его при решении других задач (см., например, замечание 4.6).

Применение Δ^2-преобразования Эйткена к последовательностям, сходящимся квадратично, эффекта ускорения не дает [176].

6.2.2. МЕТОД ВЕГСТЕЙНА*)

При выводе ***метода Вегстейна*** решения задачи о неподвижной точке (6.3) будем использовать как аналитические, так и геометрические соображения.

Пусть уже найдены: $\bar{x}_k$ — элемент строящейся здесь последовательности, и $x_{k+1} = \varphi(\bar{x}_k)$ — точка, соответствующая одному шагу МПИ, примененного к точке $\bar{x}_k$. Независимо от того, сходится начатый с $\bar{x}_k$ МПИ (рис. 6.8, где $|\xi - x_{k+1}| < |\xi - \bar{x}_k|$) или расходится (рис. 6.9 с $|\xi - x_{k+1}| > |\xi - \bar{x}_k|$), отрезок AB, параллельный оси Ox и имеющий концами точки $A(\bar{x}_k; \varphi(\bar{x}_k))$ и $B(x_{k+1}; x_{k+1})$, можно разделить точкой C так, чтобы она принадлежала вертикальной прямой $x = \xi$ (при этом во втором случае речь идет о делении отрезка внешним образом).

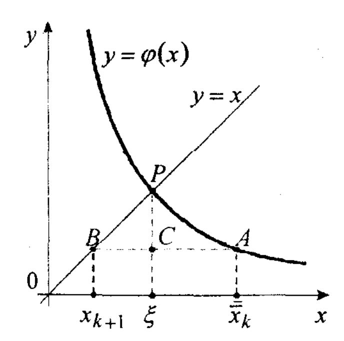


Рис. 6.8. К построению метода Вегстейна (случай сходящегося МПИ)

Рис. 6.9. К построению метода Вегстейна (случай, когда МПИ расходится)

При любых комбинациях направлений возрастания и выпуклости графика функции $y = \varphi(x)$ в окрестности неподвижной точки ξ имеет место равенство длин отрезков $BC = PC$. Различаются два случая: когда $BC = \xi - x_{k+1}$ и когда $BC = x_{k+1} - \xi$. По формуле Лагранжа соответственно имеем

$$PC = \varphi(\xi) - \varphi(\bar{x}_k) = \varphi'(\theta_k)(\xi - \bar{x}_k)$$

---

*) Ссылки на первую публикацию этого метода (1958 г) имеются в [106, 115] (в [115] его называют ***усовершенствованным методом последовательных приближений***). К сожалению, описание метода Вегстейна пока не заняло достойного места в отечественной учебной литературе по вычислительной математике. Нет упоминания о нем и в монографии [176].

или

$$PC=\varphi(\bar{x}_k)-\varphi(\xi)=\varphi'(\theta_k)(\bar{x}_k-\xi).$$

В любом случае можно утверждать, что существует точка $\theta_k\in(\bar{x}_k,\,\xi)$ или $\theta_k\in(\xi,\,\bar{x}_k)$ такая, что

$$\xi-x_{k+1}=\varphi'(\theta_k)(\xi-\bar{x}_k).$$

Разрешая это линейное уравнение относительно ξ, находим

$$\xi=x_{k+1}-\frac{x_{k+1}-\bar{x}_k}{1-\dfrac{1}{\varphi'(\theta_k)}}. \tag{6.21}$$

Если бы значение $\varphi'(\theta_k)$было известно, то тем самым задача о неподвижной точке (6.3) была бы решена точно. Заменим это неизвестное значение $\varphi'(\theta_k)$ аппроксимирующим его разностным отношением:

$$\varphi'(\theta_k)\approx\frac{\varphi(\bar{x}_k)-\varphi(\bar{x}_{k-1})}{\bar{x}_k-\bar{x}_{k-1}}=\frac{x_{k+1}-x_k}{\bar{x}_k-\bar{x}_{k-1}}.$$

Подставляя приближенное значение $\dfrac{1}{\varphi'(\theta_k)}\approx\dfrac{\bar{x}_k-\bar{x}_{k-1}}{x_{k+1}-x_k}$ в (6.21), вместо корня ξ получаем приближение к нему

$$\bar{x}_{k+1}=x_{k+1}-\frac{(x_{k+1}-x_k)(x_{k+1}-\bar{x}_k)}{(x_{k+1}-x_k)-(\bar{x}_k-\bar{x}_{k-1})}. \tag{6.22}$$

Эта итерационная формула, где $k=1,\,2,\,3,\,\ldots$, совместно с формулой

$$x_{k+1}=\varphi(\bar{x}_k)\quad(k=0,\,1,\,2,\,\ldots) \tag{6.23}$$

и начальными значениями $\bar{x}_0:=x_0$, $\bar{x}_1:=x_1$ полностью определяет метод Вегстейна для задачи (6.3).

Значение $\bar{x}_2$, получаемое по формуле Вегстейна (6.22) при заданных начальных значениях $\bar{x}_0$ и $\bar{x}_1$, совпадает со значением $\tilde{x}_2$, вычисляемым Δ^2-процессом Эйткена. Далее, т.е. при $k\ge 2$, процессы (6.20) и (6.22) различаются. Учитывая, что МПИ является составной частью метода Вегстейна, в случаях, когда $|\varphi'(x)|\le q<1$, можно заканчивать процесс вычислений, как и в методе Эйткена, по выполнении критерия (6.13).

Таким образом, для реализации метода (6.22)–(6.23) может быть предложен, например, следующий алгоритм.

Алгоритм Вегстейна

Шаг 0. Ввести x_0 (начальное приближение), $\varphi(x)$ (исходную функцию), q (оценку модуля производной), ε (допустимую абсолютную погрешность).

Шаг 1. Вычислить $x_1:=\varphi(x_0)$; положить $\bar{x}_0:=x_0$, $\bar{x}_1:=x_1$.

Шаг 2. Вычислить $x_2:=\varphi(\bar{x}_1)$.

Шаг 3. Проверить на точность: если $|x_2-\bar{x}_1|>\varepsilon(1-q)/q$, то вычислить $\bar{x}_2:=\dfrac{x_2\bar{x}_0-x_1\bar{x}_1}{x_2+\bar{x}_0-x_1-\bar{x}_1}$; переприсвоить значения $\bar{x}_0:=\bar{x}_1$, $x_1:=x_2$, $\bar{x}_1:=\bar{x}_2$ и вернуться к шагу 2.

Шаг 4. Положить $\xi:\approx x_2$ (с точностью ε).

Разумеется, проверку на точность в подобном алгоритме можно устраивать иную (что просто необходимо, если метод Вегстейна применяется в случаях, когда $|\varphi'(x)|>1$). Если нет угрозы большой потери точности из-за вычитания близких чисел, то заканчивать работу алгоритма Вегстейна лучше выводом значения $\xi:\approx\bar{x}_2$. Для вычисления значения $\bar{x}_2$ в этом алгоритме применена равносильная (6.22) формула

$$\bar{x}_{k+1}=\frac{x_{k+1}\bar{x}_{k-1}-x_k\bar{x}_k}{x_{k+1}+\bar{x}_{k-1}-x_k-\bar{x}_k},$$

имеющая несколько отличную от (6.22) структуру.

Как показывают многочисленные эксперименты с уравнениями вида $x=\varphi(x)$, особый интерес среди которых вызывают случаи, когда простые итерации дают расходящиеся последовательности (x_k), метод Вегстейна имеет определенные преимущества перед методом Эйткена по количеству обращений к вычислению значений $\varphi(x)$ для получения корня с заданной точностью. Чаще всего, метод Вегстейна еще и позволяет в более широких пределах варьировать выбор начальной точки x_0. Результаты сравнения этих двух методов на нескольких таких уравнениях представлены в табл. 6.1. В двух ее последних столбцах указано количество вычислений значений функции $\varphi(x)$ (горнеров), потребовавшееся для достижения точки $x^*\approx\xi$ (приближенного значения корня) такой, что $|x^*-\varphi(x^*)|<10^{-7}$. (Приведенные данные получены Ковалевым П.В. на IBM PC/AT-286).

Таблица 6.1

Примеры применения алгоритмов Эйткена и Вегстейна к уравнениям вида $x = \varphi(x)$

Функция $\varphi(x)$	Корень ξ (точность 10^{-6})	Значение производной $\varphi'(\xi)$ (точность 10^{-2})	Начальное приближение x_0	Число горнеров	
				метод Эйткена	метод Вегстейна
x^3+2	-1.521379	6.94	-2	18	8
			-0.5	16	9
			1	190	10
$(1-x^2)^2$	0.524889	-1.52	0	6	5
			1.3	8	5
	1.490216	7.28	1.7	20	8
			4	расходится	14
$\dfrac{1-x}{x}$	0.6180341	-2.62	0.8	10	7
	-1.618034	-0.38	-3	6	5
			-0.1	10	8
$\dfrac{1}{x}$	1	-1	1.2	6	5
			10	14	10
$e^{x \ln x}$	1	1	0.2	20	16
			1.2	20	15

Обратим внимание, что лишь для одного корня из семи в данной таблице можно говорить о сходимости МПИ.

6.3. НЕЛИНЕЙНЫЕ УРАВНЕНИЯ С ПАРАМЕТРОМ. БИФУРКАЦИИ*)

Материал § 6.1 можно интерпретировать так: нелинейная непрерывная математическая модель (6.3) некоего явления изучается путем построения и исследования соответствующей дискретной модели (6.1). Связь между этими моделями на отрезке $[a, b]$ устанавливается при выполнении двух следующих условий:

$$\varphi(x) \in [a, b] \quad \forall x \in [a, b] \quad \text{(отображение в себя)}$$

и $\hfill (6.24)$

$$|\varphi'(x)| < 1 \quad \forall x \in (a, b) \quad \text{(сжатие)}.$$

*) Сведения, излагаемые в этом пункте, имеют ознакомительный характер и опираются на статью [8]. Из других литературных источников, посвященных рассматриваемым здесь вопросам, отметим еще книгу [97, 193].

А именно, согласно теореме 6.1, эти условия являются достаточными для существования и единственности на $[a, b]$ решения ξ непрерывной задачи (6.3), причем оно может быть получено как предел последовательности (x_k) (т.е. как решение дискретной задачи (6.1)), начинающейся с любой точки $x_0 \in [a, b]$. Последнее можно расценить как устойчивость в данном смысле решения ξ дискретной модели (6.1).

Продолжим изучение взаимосвязи непрерывной и дискретной одномерных нелинейных моделей в следующем русле.

Во-первых, возьмем за основу и будем рассматривать некоторую конкретную дискретную модель, а для ее исследования привлечем соответствующую непрерывную модель.

Во-вторых, попытаемся выяснить, к чему может привести нарушение условий сходимости МПИ, т.е. условий (6.24), применительно к данной модели, и какие «тайны» могут скрываться за термином *нелинейность*.

Преследуя эти цели, введем в дискретное (6.1) и непрерывное (6.3) уравнения вещественный параметр λ, т.е. будем изучать связь между моделями вида

$$x_{k+1} = \varphi(x_k, \lambda), \quad k = 0, 1, 2, \ldots; \quad x_0 \in [a, b] \qquad (6.25)$$

и

$$x = \varphi(x, \lambda), \quad x \in [a, b]. \qquad (6.26)$$

Заметим, что не так редки ситуации, когда в приложениях математики первичными являются именно дискретные модели, а их непрерывные аналоги нужны для того, чтобы воспользоваться хорошо развитой теорией математического анализа и плодами вычислительной математики, которая, в основном, построена по принципу «непрерывная задача → дискретная аппроксимация».

Представим себе следующую весьма идеализированную картину. Пусть на некоторой ограниченной территории, например, на острове, может прокормиться не более N животных определенного вида, и пусть в начальный момент наблюдений за ними их количество было $g_0 \in (0, N)$. Будем считать, что животные ежегодно приносят потомство, и скорость размножения характеризуется некоторым параметром $\alpha > 0$. Тогда, если через g_k обозначить численность животных в k-й год после начала наблюдения, то можно предположить, что закон ежегодного изменения численности популяции грубо описывается моделью

$$g_{k+1} = \alpha g_k (N - g_k), \quad k = 0, 1, 2, \ldots . \qquad (6.27)$$

В пользу принятия такой модели говорят следующие рассуждения. Если значение g_0 начальной численности мало, то второй сомножитель в начале процесса почти постоянен, и все

зависит от коэффициента роста α: при малых α, т.е. при низкой скорости размножения, численность животных будет снижаться и, в конце концов, популяция гибнет. Если же $g_k \in [0, N]$ возрастает и приближается к максимально возможному значению N, то за счет близости к нулю второго сомножителя численность популяции естественно начнет снижаться.

Чтобы облегчить исследование модели (6.27), упростим ее заменой переменных. Переписав (6.27) в виде

$$\frac{g_{k+1}}{N} = N\alpha \cdot \frac{g_k}{N}\left(1 - \frac{g_k}{N}\right)$$

и положив $x_k := \frac{g_k}{N}$, $\lambda := N\alpha$, приходим к уравнению*)

$$x_{k+1} = \lambda x_k (1 - x_k), \tag{6.28}$$

где $k = 0, 1, 2, \ldots$, а значения x_k в соответствии со смыслом задачи должны принадлежать отрезку $[0, 1]$.

На равенство (6.28) можно смотреть как на МПИ (6.25), применяемый к задаче о неподвижной точке вида (6.26), т.е. к задаче о корнях уравнения

$$x = \lambda x (1 - x) \tag{6.29}$$

в области $x \in [0, 1]$.

Условившись не использовать далее в обозначениях функции ее явную зависимость от параметра λ, положим

$$\varphi(x) := \lambda x\,(1 - x)$$

и преобразуем эту квадратичную функцию к виду

$$\varphi(x) = \frac{\lambda}{4} - \lambda\left(x - \frac{1}{2}\right)^2.$$

Из последнего следует, что $\max \varphi(x) = \frac{\lambda}{4}$ при $x = \frac{1}{2}$ и что $\varphi(x)$ отображает отрезок $[0, 1]$ в $\left[0, \frac{\lambda}{4}\right]$. Значит, при $\lambda \in [0, 4]$ функция $\varphi(x)$ осуществляет на отрезке $[0, 1]$ отображение в себя,

---

*) Это конкретное уравнение называют ***логистическим*** [193].

т.е. при этих λ элементы последовательности (x_k), получаемой с помощью равенства (6.28), при любом $x_0 \in [0, 1]$ не выйдут за пределы $[0, 1]$, другими словами, определены при любом $k = 0, 1, 2, \ldots$ *).

Для производной данной функции $\varphi(x)$ имеем:

$$\varphi'(x) = \lambda(1 - 2x) \quad \text{и} \quad \max_{x \in [0, 1]} |\varphi'(x)| = \lambda.$$

Следовательно, при $\lambda < 1$ отображение $\varphi(x)$ является сжимающим на $[0, 1]$ и имеет единственную неподвижную точку $\xi_1 \in [0, 1]$, а именно $\xi_1 = 0$, которая является пределом последовательности (x_k) при любом начальном значении $x_0 \in [0, 1]$ (популяция гибнет по причине недостаточной скорости воспроизводства). Геометрическая иллюстрация этого случая, соответствующего выполнению условий теоремы 6.1, показана на рис. 6.10.

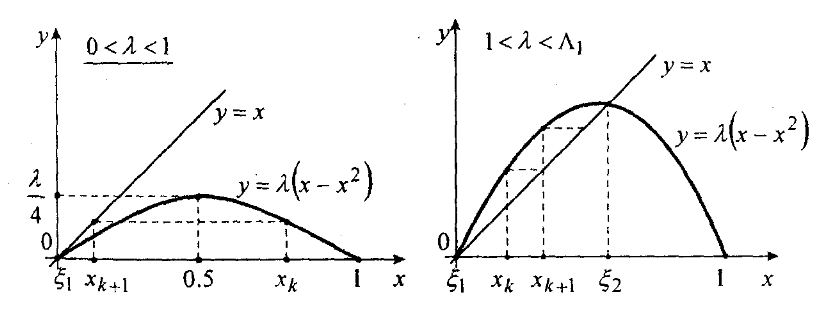


Рис. 6.10. Сходимость МПИ (6.28) к корню $\xi_1 = 0$ логистического уравнения (6.29) при $\lambda \in (0, 1)$

Рис. 6.11. Сходимость МПИ (6.28) к корню $\xi_2 = \frac{\lambda - 1}{\lambda}$ при $\lambda \in (1, \Lambda_1)$

При $1 \le \lambda \le 4$ нарушается одно из условий сходимости МПИ: $\varphi(x)$ не является функцией сжатия. Что же это влечет?

---

*) При $\lambda > 4$ нарушается соответствие между дискретной (6.25) и непрерывной (6.26) моделями. Например, при $x_0 = \frac{1}{2}$ уже $x_1 = \frac{\lambda}{4} > 1$, т.е. не принадлежит множеству, на котором определена функция $\varphi(x)$.

Очевидно, уравнение (6.29) по-прежнему сохраняет решение $\xi_1 = 0 \in [0, 1]$. Но при переходе λ через 1 это решение теряет устойчивость и появляется второе решение $\xi_2 = \frac{\lambda - 1}{\lambda} \in [0, 1]$, которое следует считать устойчивым, поскольку теперь именно к нему будет сходиться любая последовательность, определяемая начатым с $x_0 \in [0, 1]$ МПИ (6.28) (рис. 6.11). Произошло явление, которое носит название ***бифуркация решений***: вместо одного решения на рассматриваемом промежутке стало два решения*).

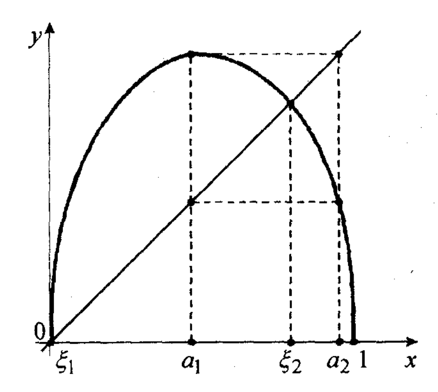


Рис. 6.12. Предельное поведение последовательности $x_{k+1} = \varphi(x_k)$, *где* $\varphi(x) := \lambda(x - x^2)$ *при* $\lambda \in (\Lambda_1, \Lambda_2)$

Сходимость (x_k) к ξ_2 будет наблюдаться не при всех $\lambda \in [1, 4]$. Оказывается, существует число $\Lambda_1 > 1$ такое, при переходе λ через которое начнет происходить ***зацикливание*** последовательности (x_k). А именно, какое бы ни взяли $x_0 \in (0, 1)$, начатая с него и продолжаемая по формуле (6.28) последовательность будет обладать тем свойством, что все ее четные члены будут иметь пределом одно число, а нечетные — другое. Это означает, что найдутся числа $a_1, a_2 \in (0, 1)$ (при каждом λ свои) такие, что $a_2 = \varphi(a_1)$ и $a_1 = \varphi(a_2)$, причем $a_1 \neq a_2 \neq \xi_i$ $(i = 1, 2)$ (см. рис. 6.12). В этом случае говорят, что дискретное отображение (6.28) имеет ***устойчивый цикл*** *периода* 2 и обозначают

*) *Bifurcus* (лат.) — раздвоенный. Термин введен К. Якоби в 1834 году. Теория бифуркаций заложена Анри Пуанкаре в конце XIX века.

его S^2. Относя это к исходной модельной задаче с животными, можно сказать, что при значениях $\lambda > \Lambda_1$ численность популяции будет меняться периодически с периодом в два «года».

Зная ситуацию качественно, нетрудно найти точно пороговое значение Λ_1, при котором появляется устойчивый цикл S^2.

Действительно, если известно, что на $(0, 1)$ имеются точки a_1, a_2 такие, что $a_2 = \varphi(a_1)$, $a_1 = \varphi(a_2)$, то значит, $a_2 = \varphi(\varphi(a_2))$, $a_1 = \varphi(\varphi(a_1))$, т.е. a_1 и a_2 — неподвижные точки отображения $\varphi^{②}(x) := \varphi(\varphi(x))$, иначе, — корни уравнения $x = \varphi(\varphi(x))$.

Так как эта суперпозиция сохраняет старые неподвижные точки $\xi_1 = 0$ и $\xi_2 = (\lambda - 1)/\lambda$, то уравнение

$$x = \lambda(\lambda x(1-x))(1 - \lambda x(1-x))$$

должно иметь четыре корня, из которых два известны (рис. 6.13).

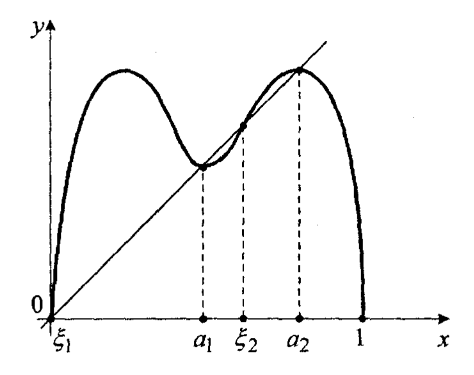


Рис. 6.13. График функции $\varphi^{②}(x) := \varphi(\varphi(x))$ *при* $\varphi(x) := \lambda(x - x^2)$, $\lambda \in (\Lambda_1, \Lambda_2)$ *и ее неподвижные точки* (ξ_1, ξ_2, a_1, a_2)

Исключив из этого уравнения известные корни, приходим к квадратному уравнению

$$x^2 - \frac{\lambda+1}{\lambda}x + \frac{\lambda+1}{\lambda^2} = 0 .$$

Положительное значение λ, которое служит границей области положительности дискриминанта $D = (\lambda^2 - 2\lambda - 3)/\lambda^2$ этого уравнения, как раз и есть искомое значение Λ_1, начиная с

которого появляются новые устойчивые неподвижные точки a_1, a_2, т.е. цикл S^2. Очевидно, это $\Lambda_1 = 3$. Устойчивость неподвижных для $\varphi^{②}(x)$ точек $a_{1,2} = \left(\lambda + 1 \pm \sqrt{\lambda^2 - 2\lambda - 3}\right)/2\lambda$ в том смысле, что они становятся точками четно-нечетного притяжения для последовательности (x_k), устанавливается непосредственной проверкой условия $\left| \frac{d}{dx} \varphi^{②}(x) \Big|_{x=a_{1,2}} \right| < 1$.

Дальнейшее увеличение λ в дискретном уравнении (6.28) ведет к тому, что начиная с некоторого Λ_2, зацикливание будет иметь более сложный характер: при каждом $\lambda \in (\Lambda_2, \Lambda_3)$, где $\Lambda_2 > 3$, $\Lambda_3 < 4$, найдутся числа a_1, a_2, a_3, a_4 (зависящие от λ) такие, что $a_2 = \varphi(a_1)$, $a_3 = \varphi(a_2)$, $a_4 = \varphi(a_3)$, $a_1 = \varphi(a_4)$, и члены последовательности (x_k) будут поочередно все сильнее притягиваться к этим числам, с какого $x_0 \in (0, 1)$ ни начинался бы процесс (6.28). Говорят, что в этом случае имеет место устойчивый цикл периода 4, т.е. цикл S^4.

Такой процесс образования новых циклов происходит с увеличением параметра λ и далее. Точки Λ_1, Λ_2, Λ_3, ..., в которых имеет место зарождение циклов S^2, S^4, S^8, ..., называются ***точками бифуркации удвоения периода***.

Этому процессу бифуркаций удвоения периода можно придать наглядный вид, если отобразить на графике зависимость ***элементов цикла*** — значений устойчивых неподвижных точек отображений $\varphi^{ⓝ}(x)$ (иначе, точек притяжения подпоследовательностей получаемой посредством МПИ (6.28) последовательности (x_k)) — от значений параметра λ при $\lambda > 1$ (см. рис. 6.14).

Последовательность (Λ_n) точек бифуркации удвоения периода обладает определенной закономерностью:

$$\frac{\Lambda_n - \Lambda_{n-1}}{\Lambda_{n+1} - \Lambda_n} \xrightarrow[n\to\infty]{} \delta = 4.66920\ldots .$$

Имеет постоянный предел, равный величине $\alpha = 2.50290\ldots$, также отношение d_{n-1}/d_n, где через d_n обозначено расстояние от точки $x = 0.5$ до ближайшего элемента цикла $S^{2^{n-1}}$, соответствующего такому значению λ, при котором $x = 0.5$ является элементом того же цикла (рис. 6.14). Числа δ и α называют ***постоянными Фейгенбаума*** в честь открывшего эти закономерности американского математика (1978 г.).

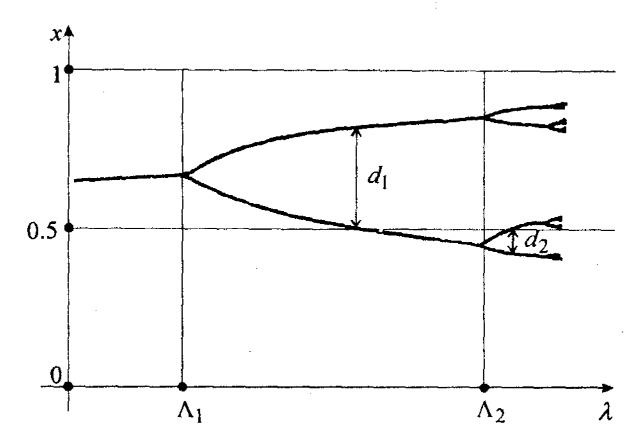


Рис. 6.14. Бифуркационная диаграмма циклов периода 2^n

Верхней границей значений λ, при которых получаемая с помощью (6.28) последовательность (x_k) ведет себя указанным образом, т.е. имеет циклы S^{2^n}, является значение $\lambda^* \approx 3.57$. Дальнейшее увеличение λ приводит к срыву цикличности. В каком-то диапазоне значений $\lambda > \lambda^*$ будет наблюдаться бесконечное хаотическое блуждание точек последовательности (x_k) в пределах промежутка $(0, 1)$, с какого бы $x_0 \in (0, 1)$ она не начиналась. Затем снова из хаоса возникают устойчивые циклы, происходят бифуркации удвоения периода и опять срыв в хаос. Такие чередования циклического (с разными периодами, например, $12 \cdot 2^n$, $10 \cdot 2^n$, $6 \cdot 2^n$, $8 \cdot 2^n$, $7 \cdot 2^n$, $5 \cdot 2^n$ и др.) и хаотического поведения последовательности (x_k) имеют место в процессе увеличения λ почти вплоть до предельного значения $\lambda = 4$. При этом самый большой (по λ) промежуток цикличности после циклов вида S^{2^n} будет иметь цикл периода $3 \cdot 2^n$ (при $\lambda \gtrsim 3.829$), играющий особую роль в теории бифуркаций.

Как зарождается порядок в хаосе, каковы связи между циклами разных периодов, какую роль играет последовательность обхода элементов цикла, что можно сказать об устойчивости тех или иных циклов — эти и другие вопросы возникают перед математиками, изучающими нелинейные отображения. Непростые ответы на них, достаточно наглядные в одномерном случае, позволяют понять природу многих сложных явлений (например, оценить принципиальные возможности долгосрочного прогнозирования погоды).

6.4. О МЕТОДАХ РЕШЕНИЯ АЛГЕБРАИЧЕСКИХ УРАВНЕНИЙ. МЕТОД БЕРНУЛЛИ

Пусть требуется найти один, несколько или все корни многочлена с действительными коэффициентами

$$P_n(x) := a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n, \qquad (6.30)$$

т.е. решить в каком-то указанном смысле уравнение

$$P_n(x) = 0. \qquad (6.31)$$

Если речь идет о нахождении только действительных корней и, особенно, если нужно найти не все, а только некоторые из них, то есть резон в применении к алгебраическому уравнению (6.31) какого-либо из рассмотренных выше методов решения нелинейных скалярных уравнений $f(x)=0$ с $P_n(x)$ в роли $f(x)$. При этом следует обратить внимание на то, что $P_n(x)$ является определенной на всей действительной оси бесконечно дифференцируемой функцией, и здесь можно применять методы, использующие производные $P_n(x)$. Поскольку дифференцирование понижает степень многочлена, вычисление значений производных потребует даже меньше арифметических действий, чем вычисление значений исходного многочлена. Отсюда — целесообразность нахождения отдельных корней многочлена (в том числе и комплексных, см. [61, 72, 115, 129, 158]) методом Ньютона или гибридным алгоритмом, его использующим.

Решение алгебраических уравнений (6.31) любым итерационным способом требует многократного вычисления значений многочленов (6.30). Эта промежуточная задача для многочленов намного проще, чем задача вычисления значений трансцендентных функций: нужно всего лишь при заданном $x = x_0$ простым перемножением находить степени x_0^i при $i = 1, 2, \ldots, n$ (это $n-1$ умножение), затем умножить их на коэффициенты (n умножений) и сложить результаты (n сложений). Однако в практике вычислений используют более эффективный способ вычисления значений многочленов — ***схему Горнера***, позволяющую почти вдвое уменьшить количество умножений*). Выведем этот способ.

---

*) Такой способ был известен в Китае еще в средние века и назывался ***Тянь-юань***, а затем в начале XIX века был «переоткрыт» в Европе англичанином Горнером и итальянцем Руффини.

Согласно теореме Безу, при любом x_0 найдутся числа b_i ($i = 0, 1, 2, \ldots, n$) такие, что

$$a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n \equiv$$
$$\equiv (x - x_0)\left(b_0x^{n-1} + b_1x^{n-2} + \ldots + b_{n-2}x + b_{n-1}\right) + b_n,$$

причем $b_n = P_n(x_0)$. Выполняя умножение в правой части тождества и приравнивая коэффициенты при одинаковых степенях x, получим совокупность $n+1$ равенств:

$$b_0 = a_0, \quad b_1 = a_1 + x_0b_0, \quad b_2 = a_2 + x_0b_1, \quad \ldots, \quad b_n = a_n + x_0b_{n-1}.$$

Таким образом, после n умножений (на одно и то же число x_0) и n сложений приходим к искомому результату $b_n = P_n(x_0)$.

Однотипность вычислений, производимых в схеме Горнера, позволяет организовать их в цикл, определяемый формулой

$$b_i = a_i + x_0b_{i-1},$$

где $i = 1, 2, \ldots, n$, $b_0 := a_0$; на его выходе имеем значение $P_n(x_0) := b_n$.

Числа $b_0, b_1, \ldots, b_n$ далее будем называть ***коэффициентами схемы Горнера***. При ручном счете эти коэффициенты удобно записывать в процессе вычислений под соответствующими размещенными в один ряд коэффициентами данного многочлена (6.30) в виде следующей таблицы:

	a_0	a_1	a_2	$\ldots$	a_{n-1}	a_n
x_0	b_0	b_1	b_2	$\ldots$	b_{n-1}	b_n

Пример 6.1. Доказать, что число 2 является единственным рациональным корнем многочлена $P_5(x) := x^5 - x^4 - 3x^3 + 2x + 4$, причем простым.

Все доказательство можно отразить следующей таблицей:

	1	−1	−3	0	2	4
2	1	1	−1	−2	−2	0
1	1	2	1	−1	−3	
−1	1	0	−1	−1	−1	
2	1	3	5	8	14	
−2	1	−1	1	−4	6	

В ее первой строке находятся коэффициенты данного многочлена $P_5(x)$, во второй строке — коэффициенты схемы Горнера, примененной к $P_5(x)$ при $x_0 = 2$, означающие, что $P_5(2) = 0$ и $P_5(x) = (x-2)P_4(x)$, где $P_4(x) = x^4 + x^3 - x^2 - 2x - 2$ (см. тождество, лежащее в основе вывода

схемы Горнера). Остальные строки — результаты применения схемы Горнера к многочлену $P_4(x)$ (что отражено в таблице подчеркиванием) в точках $\pm 1, \pm 2$, являющихся делителями свободного члена многочлена с целыми коэффициентами $P_4(x)$ со старшим коэффициентом 1. Так как $P_4(1)=-3$, $P_4(-1)=-1$, $P_4(2)=14$, и $P_4(-2)=6$, то $P_4(x)$ не имеет рациональных корней, а значит, и $P_5(x)$ не имеет других рациональных корней, кроме 2, причем 2 не может быть двукратным корнем.

Как видно из приведенного примера, схему Горнера удобно использовать для понижения степени алгебраического уравнения выделением линейного множителя, соответствующего известному вещественному корню. Поскольку в подавляющем большинстве случаев эти корни бывают известны лишь приближенно, при понижении степени алгебраического уравнения неизбежна потеря точности, с которой могут быть найдены последующие корни. Это обстоятельство заставляет с осторожностью относиться к выигрышу в вычислительных затратах, достигаемому таким понижением степени, когда находятся несколько корней многочлена последовательно корень за корнем.

Существует способ вычисления корня многочлена (6.30) последовательно цифра за цифрой непосредственно по схеме Горнера, применяемой к специальным образом преобразованному многочлену; в [61] такой способ называется ***методом Горнера***.

Более распространен метод выделения множителей*) [20, 72, 99, 102, 129]. Последовательное выделение линейного множителя этим методом базируется на применении схемы Горнера в точках последовательности приближений $x_0, x_1, x_2, \ldots$ к искомому корню ξ многочлена $P_n(x)$. Это оказывается равнозначным простым итерациям вида

$$x_{k+1} = \frac{P_n(0)x_k}{P_n(0)-P_n(x_k)}, \quad k=0,1,2,\ldots,$$

что позволяет использовать известные критерии сходимости МПИ. При нахождении пары комплексно сопряженных корней выделяют к в а д р а т и ч н ы й множитель. Для этих целей нужна эффективная схема деления с остатком многочлена на квадратный трехчлен, которую нетрудно получить по аналогии с выводом схемы Горнера.

В случаях, когда вычисление корней многочлена ориентировано на применение итерационных методов решения нелинейных уравнений общего вида, встает вопрос о сужении области

---

*) На самом деле, имеется несколько таких методов, из которых наиболее известен ***метод Лина*** (***предпоследнего остатка***), который и имеется здесь в виду.

поиска корней. Здесь опять может оказаться полезной схема Горнера. Справедливо утверждение:

если все коэффициенты схемы Горнера, примененной к многочлену $P_n(x)$ в точке $x=\beta>0$, положительны, то правее β на оси Ox действительных корней $P_n(x)$ нет.

Это следует из тождества

$$P_n(x) \equiv (x-\beta)\left(b_0 x^{n-1} + b_1 x^{n-2} + \ldots + b_{n-2}x + b_{n-1}\right) + b_n,$$

показывающего, что $P_n(x)>0$ для любых $x \ge \beta > 0$, в силу положительности всех b_i $(i=0,1,2,\ldots,n)$ и разности $x-\beta$.

Так, в примере 6.1, глядя на приведенную там таблицу, можно сказать, что многочлен $P_4(x)$ (а значит, и $P_5(x)$) не имеет действительных корней, больших 2.

Есть более конструктивные способы указания границ всех корней многочлена.

Например, можно утверждать, что

за верхнюю границу положительных корней многочлена $P_n(x)$ с $a_0>0$ можно принять число

$$R = 1 + \sqrt[m]{\frac{B}{a_0}}, \tag{6.32}$$

где m — индекс первого отрицательного коэффициента в ряду $a_1, a_2, \ldots, a_n$ (иначе, разность между показателем степени многочлена и показателем степени первого отрицательного члена), *а B — максимум модулей всех отрицательных коэффициентов**).*

Действительно, заменив в $P_n(x)$ неотрицательные коэффициенты $a_1, a_2, \ldots, a_{m-1}$ нулями, а все последующие — числом $-B$, при $x>1$ имеем:

$$P_n(x) \ge a_0x^n - B\left(x^{n-m} + \ldots + x + 1\right) =$$

$$= a_0x^n - B\frac{x^{n-m+1}-1}{x-1} > a_0x^n - \frac{Bx^{n-m+1}}{x-1} =$$

$$= \frac{x^{n-m+1}}{x-1}\left[a_0x^{m-1}(x-1) - B\right] > \frac{x^{n-m+1}}{x-1}\left[a_0(x-1)^m - B\right].$$

---

*) В одних литературных источниках такое нахождение правой границы действительных корней называют ***методом Лагранжа*** [46, 61], в других — ***методом Маклорена*** [72].

Так как последнее выражение в этой цепочке равенств и неравенств равно нулю при $x = R$ и больше нуля при $x > R$, значит, $P_n(x) > 0$ при всех $x \geq R$, т.е. правее R действительных корней нет.

Любой метод нахождения верхней границы положительных корней можно приспособить для нахождения нижней (левой) границы отрицательных корней. Для этого достаточно преобразовать многочлен $P_n(x)$ заменой $t = -x$ и для положительных корней многочлена $(-1)^n P_n(t)$ найти верхнюю границу R^*(например, по формуле (6.32)); тогда число $-R^*$ будет искомой нижней границей*). Так же, используя известный метод нахождения верхней границы в многочленах $P_n\left(\frac{1}{x}\right)$ и $(-1)^n P_n\left(-\frac{1}{x}\right)$, можно найти нижнюю границу положительных и верхнюю границу отрицательных корней многочлена $P_n(x)$, т.е. отделить его корни от нуля.

Зачастую более хорошие результаты показывает ***метод Вестерфильда*** получения симметричных границ расположения всех корней многочлена [72]. Согласно ему,

> *модули всех корней* (в том числе и комплексных) *приведенного многочлена* $P_n(x)$ (т.е. при $a_0 = 1$) *лежат в круге, радиус которого не превосходит суммы двух наибольших из чисел* $\sqrt[m]{|a_m|}$ *, где* $m = 1, 2, \ldots, n$.

Решение проблемы изоляции корней алгебраического уравнения не имеет особой специфики, выделяющей ее из более общего случая нелинейных скалярных уравнений. Более продвинутым здесь является решение вопроса о количестве действительных корней. Можно указать простые способы выяснения в о з м о ж н о г о количества положительных и отрицательных корней по числу перемен знаков в последовательностях коэффициентов $P_n(x)$ и $P_n(-x)$.

Так, ***теорема Декарта*** говорит о том, что

> *число положительных корней уравнения* (6.31) *с учетом их кратностей равно числу перемен знаков в последовательности коэффициентов* $a_0, a_1, \ldots, a_n$ (без учета нулевых коэффициентов) *или на четное число меньше* [61].

Более точный ответ на вопрос о числе действительных корней алгебраического уравнения можно получить с помощью широко известной в алгебре ***теоремы Штурма*** (см. [20, 46, 61] и др.).

*) Множитель $(-1)^n$ поставлен ради положительности старшего коэффициента преобразованного многочлена.

Если при этом уже затрачены усилия на составление системы Штурма, то ее можно использовать и для нахождения промежутков изоляции действительных корней.

Имеются и другие способы нахождения границ действительных и комплексных корней алгебраических уравнений, выяснения количества положительных и отрицательных корней, а также их изоляции.

Одним из наиболее эффективных методов нахождения в с е х или почти всех корней алгебраического уравнения, как вещественных, так и комплексных, является ***метод Лобачевского***, предложенный выдающимся русским математиком в 1834 году*). Основная идея метода заключается в последовательном применении операции квадрирования корней. Суть ее такова.

С помощью обобщенной теоремы Виета легко показать, что если корни $\xi_1, \xi_2, \ldots, \xi_n$ уравнения (6.31) сильно отличатся по модулю (что, кстати, гарантирует их вещественность), а именно, $|\xi_1| >> |\xi_2| >> \cdots >> |\xi_n|$, то

$$\xi_1 \approx -\frac{a_1}{a_0}, \quad \xi_2 \approx -\frac{a_2}{a_1}, \quad \ldots \quad , \quad \xi_n \approx -\frac{a_n}{a_{n-1}}.$$

При возведении их в натуральную степень будут получаться все более удаленные друг от друга числа. Одна из макроопераций метода Лобачевского (***квадрирование корней***) состоит в том, что от данного уравнения с корнями $\xi_1, \xi_2, \ldots, \xi_n$ переходят к новому уравнению той же степени с коэффициентами

$$A_i = a_i^2 + 2\sum_{j=1}^{i}(-1)^j a_{i-j}a_{i+j} \quad (i = 0, 1, 2, \ldots, n) \qquad (6.33)$$

и корнями $\mu_i = -\xi_i^2$ $(i = 1, 2, \ldots, n)$. После ее многократного применения, когда в (6.33) будет практически сведена на нет роль второго слагаемого (удвоенной суммы парных произведений), извлечением корней соответствующих степеней можно приближенно найти модули корней исходного уравнения.

Более подробно о методе Лобачевского, его реализациях, разных ситуациях, с которыми можно встретиться при его применении, см. в книгах [20, 61, 72, 98, 129]. Одна из последних

*) Этот метод называют также ***методом Лобачевского-Греффе*** или ***методом Данделена*** в честь швейцарского математика Греффе и французского математика Данделена, причастных к одним из первых версий метода. Впоследствии метод Лобачевского неоднократно совершенствовался.

версий этого не самого простого метода содержится в брошюре А.А. Беланова [17] (ориентированной, правда, на программируемые микрокалькуляторы, а не на компьютеры).

Рассмотрим простой способ приближенного вычисления наибольшего по модулю действительного корня алгебраического уравнения (6.31), который был опубликован И. Бернулли в 1732 г. и называется ***методом Бернулли***.

Запишем уравнение (6.31) в виде

$$x^n = c_1 x^{n-1} + c_2 x^{n-2} + \ldots + c_{n-1} x + c_n.$$

С помощью коэффициентов $c_i = -\frac{a_i}{a_0}$ $(i = 1, 2, \ldots, n)$ этого уравнения будем строить последовательность $(u_{n+k})_{k=1}^{\infty}$ по рекуррентной формуле

$$u_{n+k} = c_1 u_{n+k-1} + c_2 u_{n+k-2} + \ldots + c_{n-1} u_{k+1} + c_n u_k, \qquad (6.34)$$

начиная этот процесс при $k = 1, 2, \ldots, n$ со значений

$$u_1 = 0, \quad u_2 = 0, \quad \ldots, \quad u_{n-1} = 0, \quad u_n = k$$

(как это рекомендовано Хильдебрандом [72]).

На основе обобщенной теоремы Виета можно выяснить, что эта последовательность обладает замечательным свойством: если $\xi_1, \xi_2, \ldots, \xi_n$ — корни многочлена $P_n(x)$, то

$$\xi_1 + \xi_2 + \ldots + \xi_n = u_{n+1},$$

$$\xi_1^2 + \xi_2^2 + \ldots + \xi_n^2 = u_{n+2},$$

. . . .

$$\xi_1^k + \xi_2^k + \ldots + \xi_n^k = u_{n+k}$$

. . . .

Взяв отношение двух соседних членов последовательности (u_{n+k}), выраженных через степени корней, имеем:

$$\frac{u_{n+k+1}}{u_{n+k}} = \frac{\xi_1^{k+1} + \xi_2^{k+1} + \ldots + \xi_n^{k+1}}{\xi_1^k + \xi_2^k + \ldots + \xi_n^k} = \frac{\xi_1^{k+1}\left[1 + \left(\frac{\xi_2}{\xi_1}\right)^{k+1} + \ldots + \left(\frac{\xi_n}{\xi_1}\right)^{k+1}\right]}{\xi_1^k\left[1 + \left(\frac{\xi_2}{\xi_1}\right)^k + \ldots + \left(\frac{\xi_n}{\xi_1}\right)^k\right]}. \qquad (6.35)$$

Если $|\xi_1| > |\xi_i| \quad \forall i \in \{2, \ldots, n\}$, то, очевидно, $\frac{u_{n+k+1}}{u_{n+k}} \underset{k\to\infty}{\to} \xi_1$.
К такому же результату придем и в случае, когда ξ_1 — не простой, а m-кратный корень (выражения в квадратных скобках в числителе и в знаменателе (6.35) имеют пределом число m).

Если сходимость последовательности отношений $\frac{u_{n+k+1}}{u_{n+k}}$ не обнаруживается, но существует предел последовательности величин

$$\frac{u_{n+k+2}}{u_{n+k}} = \xi_1^2 \frac{1 + \left(\frac{\xi_2}{\xi_1}\right)^{k+2} + \ldots + \left(\frac{\xi_n}{\xi_1}\right)^{k+2}}{1 + \left(\frac{\xi_2}{\xi_1}\right)^k + \ldots + \left(\frac{\xi_n}{\xi_1}\right)^k},$$

то в этом случае данное уравнение имеет два действительных корня $\xi_1 = -\xi_2$, наибольших по модулю. При достаточно больших k (желательно, четных) они могут быть найдены приближенным равенством $\xi_{1,2} \approx \pm\sqrt{\frac{u_{n+k+2}}{u_{n+k}}}$. Хаотическое поведение последовательности отношений $\frac{u_{n+k+2}}{u_{n+k}}$ говорит о том, что превалирующими являются комплексные корни. Незначительной доработкой метод Бернулли можно приспособить и для этого случая (см. [72]).

Как видно, метод Бернулли весьма близок к степенному методу решения частичной проблемы собственных значений, более тщательно рассмотренному в § 4.2. Зная в тонкостях один из этих методов, можно многое сказать о поведении другого.

УПРАЖНЕНИЯ

6.1. Дайте обоснованное заключение о факте и скорости сходимости последовательности

$$x_{k+1} = \frac{1}{4}x_k^4 - \frac{1}{2}x_k + 1, \quad k = 0, 1, 2, \ldots, \quad x_0 = 0$$

на отрезке [0, 1]. С какой точностью можно приблизиться к $\lim\limits_{k\to\infty} x_k$ за 10 шагов? Запишите итерационный процесс Ньютона, имеющий тот же предел.

6.2. Запишите сходящийся процесс простых итераций для нахождения корня уравнения $x^3 - 2x^2 - 4x - 7 = 0$, изолированного на промежутке $[3, 4]$.

6.3. Подготовьте расчетные формулы для нахождения корня уравнения $2 - \lg x - x = 0$ методом простых итераций. Выбрав начальное приближение, подсчитайте, за сколько итерационных шагов можно гарантировать получение корня с точностью $\varepsilon = 10^{-6}$. Сделайте четыре шага МПИ и по два полных шага Δ^2-процесса Эйткена и метода Вегстейна.

6.4. Сравните эффективность следующих двух подходов к вычислению значений $P_n(x_k)$ и $P_n'(x_k)$ при нахождении алгебраического корня уравнения $P_n(x) = 0$ методом Ньютона $x_{k+1} = x_k - \dfrac{P_n(x_k)}{P_n'(x_k)}$:

А) вычислив степени x_k^i, использовать их затем в $P_n(x_k)$ и $P_n'(x_k)$;

Б) применить схему Горнера сначала к многочлену $P_n(x)$ в точке x_k, затем к его неполному частному — результату первого ее применения (правомерность этого подхода обоснуйте).

6.5. Разработайте алгоритм, позволяющий найти наибольший корень многочлена $P_n(x)$ с заданной точностью только одними проверками на положительность коэффициентов схемы Горнера.

6.6. А) Докажите, что если числа R_1 и R_2 — правые границы действительных корней многочленов $Q(x)$ и $T(x)$ соответственно, то за верхнюю границу действительных корней многочлена $P(x) := Q(x) + T(x)$ можно принять число $R = \max\{R_1, R_2\}$.

Б) Для многочлена $P_5(x) = x^5 - 2x^4 + 15x^2 - 32x + 1$ найдите границы положительных корней сначала непосредственно по формуле (6.32), затем на основе результата А, выполняя подходящее представление $P_5(x)$ суммой $Q(x) + T(x)$ так, чтобы получить как можно более точную границу.

В) Найдите границу модулей всех корней данного многочлена $P_5(x)$ методом Вестерфильда.

6.7. Пусть ξ_1, ξ_2 — вещественные корни уравнения $x^2 - px + q = 0$. Убедитесь в справедливости равенств

$$\xi_1 + \xi_2 = u_3, \quad \xi_1^2 + \xi_2^2 = u_4, \quad \xi_1^3 + \xi_2^3 = u_5,$$

где u_i — элементы последовательности Бернулли–Хильдебранда, задаваемой формулой (6.34).

6.8. Дано уравнение $x^4 - 3x^3 - 7x^2 + 15x + 18 = 0$. Сделайте 5–6 приближений к его старшему корню методом Бернулли.

ГЛАВА 7 МЕТОДЫ РЕШЕНИЯ СИСТЕМ НЕЛИНЕЙНЫХ УРАВНЕНИЙ

Рассматривается ряд методов решения систем алгебраических и трансцендентных уравнений. Среди них метод простых итераций, метод Ньютона в разных модификациях (в частности, n-полюсный метод Ньютона), метод Брауна, метод секущих Бройдена. Показывается связь между данной задачей и задачей безусловной минимизации функции нескольких переменных. Проводится сравнение методов на примере решения конкретной системы. С единых позиций изучается сходимость основного и упрощенного методов Ньютона и метода, получаемого из метода Ньютона применением итерационного процесса Шульца для приближенного обращения матриц Якоби.

7.1. ВЕКТОРНАЯ ЗАПИСЬ НЕЛИНЕЙНЫХ СИСТЕМ. МЕТОД ПРОСТЫХ ИТЕРАЦИЙ

Пусть требуется решить систему уравнений

$$\begin{cases} f_1(x_1, x_2, \ldots, x_n) = 0, \\ f_2(x_1, x_2, \ldots, x_n) = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ f_n(x_1, x_2, \ldots, x_n) = 0, \end{cases} \qquad (7.1)$$

где $f_1, f_2, \ldots, f_n$ — заданные, вообще говоря, нелинейные (среди них могут быть и линейные) вещественнозначные функции n вещественных переменных $x_1, x_2, \ldots, x_n$.

Обозначив

$$\mathbf{x} := \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad F(\mathbf{x}) := \begin{pmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_n(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} f_1(x_1, x_2, \ldots x_n) \\ f_2(x_1, x_2, \ldots x_n) \\ \vdots \\ f_n(x_1, x_2, \ldots x_n) \end{pmatrix}, \quad \mathbf{0} := \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

данную систему (7.1) можно записать одним уравнением

$$F(\mathbf{x}) = \mathbf{0} \qquad (7.1а)$$

относительно векторной функции F векторного аргумента $\mathbf{x}$. Таким образом, исходную задачу можно рассматривать как задачу о нулях нелинейного отображения $F: \mathbf{R}_n \to \mathbf{R}_n$. В этой по-

становке она является прямым обобщением основной задачи предыдущей главы – задачи построения методов нахождения нулей одномерных нелинейных отображений. Фактически это та же задача, только в пространствах большей размерности. Поэтому можно как заново строить методы ее решения на основе разработанных выше подходов, так и осуществлять формальный перенос выведенных для скалярного случая расчетных формул. В любом случае следует позаботиться о правомочности тех или иных операций над векторными переменными и векторными функциями, а также о сходимости получаемых таким способом итерационных процессов. Часто теоремы сходимости для этих процессов являются тривиальными обобщениями соответствующих результатов, полученных для методов решения скалярных уравнений. Однако не все результаты и не все методы можно перенести со случая $n=1$ на случай $n \geq 2$. Например, здесь уже не будут работать методы дихотомии, поскольку множество векторов не упорядочено. В то же время, переход от $n=1$ к $n \geq 2$ вносит в задачу нахождения нулей нелинейного отображения свою специфику, учет которой приводит к новым методам и к различным модификациям уже имеющихся. В частности, большая вариативность методов решения нелинейных систем связана с разнообразием способов, которыми можно решать линейные алгебраические задачи, возникающие при пошаговой линеаризации данной нелинейной вектор-функции $F(\mathbf{x})$.

Начнем изучение методов решения нелинейных систем с наиболее простого метода.

Пусть система (7.1) имеет вид (преобразована к виду):

$$\begin{cases} x_1 = \varphi_1(x_1, x_2, \dots, x_n), \\ x_2 = \varphi_2(x_1, x_2, \dots, x_n), \\ \dots\dots\dots\dots\dots\dots \\ x_n = \varphi_n(x_1, x_2, \dots, x_n), \end{cases} \tag{7.2}$$

или иначе, в компактной записи,

$$\mathbf{x} = \Phi(\mathbf{x}), \tag{7.2a}$$

где

$$\Phi(\mathbf{x}) := \begin{pmatrix} \varphi_1(\mathbf{x}) \\ \varphi_2(\mathbf{x}) \\ \vdots \\ \varphi_n(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \varphi_1(x_1, x_2, \dots, x_n) \\ \varphi_2(x_1, x_2, \dots, x_n) \\ \vdots \\ \varphi_n(x_1, x_2, \dots, x_n) \end{pmatrix}.$$

Для этой ***задачи о неподвижной точке*** нелинейного отображения $\Phi : \mathbf{R}_n \to \mathbf{R}_n$ запишем формально рекуррентное равенство

$$\mathbf{x}^{(k+1)} = \Phi\left(\mathbf{x}^{(k)}\right), \quad k = 0, 1, 2, \dots, \tag{7.3}$$

которое определяет ***метод простых итераций (МПИ)*** (или ***метод последовательных приближений***) для задачи (7.2).

Если начать процесс построения последовательности $\left(\mathbf{x}^{(k)}\right)$ с некоторого вектора $\mathbf{x}^{(0)} = \left(x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)}\right)^T$ и продолжить по формуле (7.3), то при определенных условиях эта последовательность со скоростью геометрической прогрессии будет приближаться к вектору $\mathbf{x}^*$ — неподвижной точке отображения $\Phi(\mathbf{x})$. А именно, справедлива следующая теорема.

Теорема 7.1. *Пусть функция $\Phi(\mathbf{x})$ и замкнутое множество $M \subseteq D(\Phi) \subseteq \mathbf{R}_n$ таковы, что:*

1) $\Phi(\mathbf{x}) \in M \quad \forall\, \mathbf{x} \in M$;
2) $\exists q < 1 : \|\Phi(\mathbf{x}) - \Phi(\tilde{\mathbf{x}})\| \leq q \cdot \|\mathbf{x} - \tilde{\mathbf{x}}\| \quad \forall\, \mathbf{x}, \tilde{\mathbf{x}} \in M$.

Тогда $\Phi(\mathbf{x})$ имеет в M единственную неподвижную точку $\mathbf{x}^*$; последовательность $\left(\mathbf{x}^{(k)}\right)$, определяемая МПИ (7.3), при любом $\mathbf{x}^{(0)} \in M$ сходится к $\mathbf{x}^*$ и справедливы оценки

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \leq \frac{q}{1-q}\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \leq \frac{q^k}{1-q}\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right\| \quad \forall\, k \in \mathbf{N}.$$

Доказательство этой теоремы почти полностью повторяет доказательство теоремы 6.1 и даже несколько проще его. Однако и практическая ценность такой теоремы не так велика из-за неконструктивности ее условий. В случаях, когда имеется хорошее начальное приближение $\mathbf{x}^{(0)}$ к решению $\mathbf{x}^*$, больший интерес для приложений может представить следующий аналог теоремы 6.2.

Теорема 7.2. *Пусть функция $\Phi(\mathbf{x})$ дифференцируема*) в замкнутом шаре**) $S\left(\mathbf{x}^{(0)}, r\right) \subseteq D(\Phi)$, причем $\exists q \in (0, 1)$:*

---

*) Здесь и далее под дифференцируемостью понимается дифференцируемость по Фреше (см. приложение 1).

**) Т.е. на множестве S точек $\mathbf{x} \in \mathbf{R}_n$ таких, что $\left\|\mathbf{x} - \mathbf{x}^{(0)}\right\| \leq r$.

$\sup\limits_{\mathbf{x}\in S}\|\Phi'(\mathbf{x})\| \le q$. *Тогда, если центр* $\mathbf{x}^{(0)}$ *и радиус* r *шара* S, *таковы, что* $\left\|\mathbf{x}^{(0)} - \Phi\left(\mathbf{x}^{(0)}\right)\right\| \le r(1-q)$, *то справедливо заключение теоремы* 7.1 *с* $M = S$.

Как видим, теорема 7.2 — это просто соответствующим образом отредактированная для многомерного случая теорема 6.2. Переложение доказательства теоремы 6.2 на этот случай менее тривиально из-за того, что в $\mathbf{R}_n$ при $n > 1$ не работает дифференциальная теорема о среднем в привычном виде формулы конечных приращений Лагранжа. Но если воспользоваться теоремой о среднем, например, в виде неравенства (см. [80, 139 и др.])

$$\|F(\mathbf{x}) - F(\widetilde{\mathbf{x}})\| \le \sup_{0<t<1}\|F'(\widetilde{\mathbf{x}} + t(\mathbf{x} - \widetilde{\mathbf{x}}))\| \cdot \|\mathbf{x} - \widetilde{\mathbf{x}}\|,$$

то доказательство теоремы 7.2 становится более прозрачным. Другой путь доказательства теоремы 7.2 — покоординатное применение одномерной формулы Лагранжа и использование конкретных матричных и векторных пар норм [44, 61].

Если потребовать непрерывную дифференцируемость $\Phi(\mathbf{x})$, то более просто перейти от теоремы 7.1 к теореме 7.2, применив следующее утверждение [177].

Лемма 7.1. *Пусть функция* $\Phi : \mathbf{R}_n \to \mathbf{R}_n$ *непрерывно дифференцируема на множестве* $M \subseteq D(\Phi)$ *и пусть* $\|\Phi'(\mathbf{x})\| \le L \quad \forall \mathbf{x} \in M$. *Тогда* $\Phi(\mathbf{x})$ *удовлетворяет на множестве* M *условию Липшица*

$$\|\Phi(\mathbf{x}) - \Phi(\widetilde{\mathbf{x}})\| \le L\|\mathbf{x} - \widetilde{\mathbf{x}}\|.$$

Запись МПИ (7.3) в развернутом виде, т.е. совокупность рекуррентных равенств

$$\begin{cases} x_1^{(k+1)} = \varphi_1\left(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}\right), \\ x_2^{(k+1)} = \varphi_2\left(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}\right), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^{(k+1)} = \varphi_n\left(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}\right), \end{cases}$$

напоминает изученный ранее (см. гл.3) МПИ для СЛАУ, который укладывается в эту схему, если все функции φ_j — линейные. Учитывая, что в линейном случае, как правило, по сравнению с МПИ более эффективен метод Зейделя (см. (3.12) и теорему 3.7), здесь тоже может оказаться полезной подобная модификация. А именно, вместо (7.3) можно реализовать следующий ***метод покоординатных итераций***:

$$\begin{cases} x_1^{(k+1)} = \varphi_1\left(x_1^{(k)}, x_2^{(k)}, \ldots, x_{n-1}^{(k)}, x_n^{(k)}\right), \\ x_2^{(k+1)} = \varphi_2\left(x_1^{(k+1)}, x_2^{(k)}, \ldots, x_{n-1}^{(k)}, x_n^{(k)}\right), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n^{(k+1)} = \varphi_n\left(x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_{n-1}^{(k+1)}, x_n^{(k)}\right). \end{cases} \qquad (7.4)$$

Заметим, что как и для линейных систем, отдельные уравнения в методе (7.4) неравноправны, т.е. перемена местами уравнений системы (7.2) может изменить в каких-то пределах число итераций и вообще ситуацию со сходимостью последовательности итераций. Чтобы применить метод простых итераций (7.3) или его зейделеву модификацию (7.4) к исходной системе (7.1), нужно, как и в скалярном случае, сначала тем или иным способом привести ее к виду (7.2). Это можно сделать, например, умножив (7.1а) на некоторую неособенную $n \times n$-матрицу $-\mathbf{A}$ и прибавив к обеим частям уравнения $-\mathbf{A}F(\mathbf{x}) = \mathbf{0}$ вектор неизвестных $\mathbf{x}$. Полученная система

$$\mathbf{x} = \mathbf{x} - \mathbf{A}F(\mathbf{x})$$

эквивалентна данной и имеет вид задачи о неподвижной точке (7.2а). Проблема теперь состоит лишь в подборе матричного параметра $\mathbf{A}$ такого, при котором вектор-функция $\Phi(\mathbf{x}) := \mathbf{x} - \mathbf{A}F(\mathbf{x})$ обладала бы нужными свойствами.

7.2. МЕТОД НЬЮТОНА, ЕГО РЕАЛИЗАЦИИ И МОДИФИКАЦИИ

Пусть $(\mathbf{A}_k)$ — некоторая последовательность невырожденных вещественных $n \times n$-матриц. Тогда, очевидно, последовательность задач

$$\mathbf{x} = \mathbf{x} - \mathbf{A}_k F(\mathbf{x}), \quad k = 0, 1, 2, \ldots$$

имеет те же решения, что и исходное уравнение (7.1а), и для приближенного нахождения этих решений можно формально записать итерационный процесс

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{A}_k F\left(\mathbf{x}^{(k)}\right), \quad k = 0, 1, 2, \ldots, \qquad (7.5)$$

имеющий вид метода простых итераций (7.3) при

$\Phi(\mathbf{x}) := \Phi_k(\mathbf{x}) := \mathbf{x} - \mathbf{A}_k F(\mathbf{x})$. В случае $\mathbf{A}_k \equiv \mathbf{A}$ это, как показано в конце предыдущего параграфа, — действительно МПИ с линейной сходимостью последовательности $\left(\mathbf{x}^{(k)}\right)$. Если же $\mathbf{A}_k$ различны при разных k, то формула (7.5) определяет большое семейство итерационных методов с матричными параметрами $\mathbf{A}_k$. Рассмотрим некоторые из методов этого семейства.

Положим $\mathbf{A}_k := \left[F'\left(\mathbf{x}^{(k)}\right)\right]^{-1}$, где

$$F'(\mathbf{x}) = J(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}$$

— матрица Якоби вектор-функции $F(\mathbf{x})$. Подставив это $\mathbf{A}_k$ в (7.5), получаем явную формулу ***метода Ньютона***

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left[F'\left(\mathbf{x}^{(k)}\right)\right]^{-1} F\left(\mathbf{x}^{(k)}\right), \qquad (7.6)$$

обобщающего на многомерный случай скалярный метод Ньютона (5.14). Эту формулу, требующую обращения матриц на каждой итерации, можно переписать в неявном виде:

$$F'\left(\mathbf{x}^{(k)}\right)\left(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right) = -F\left(\mathbf{x}^{(k)}\right). \qquad (7.7)$$

Применение (7.7) предполагает при каждом $k = 0, 1, 2, \ldots$ решение линейной алгебраической системы

$$F'\left(\mathbf{x}^{(k)}\right)\mathbf{p}^{(k)} = -F\left(\mathbf{x}^{(k)}\right)$$

относительно векторной ***поправки*** $\mathbf{p}^{(k)} = \left(p_1^{(k)}, p_2^{(k)}, \ldots, p_n^{(k)}\right)^T$, а затем прибавление этой поправки к текущему приближению для получения следующего:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{p}^{(k)}.$$

К решению таких линейных систем можно привлекать самые разные методы как прямые, так и итерационные (см. гл. 2, 3) в зависимости от размерности n решаемой задачи и специфики матриц Якоби $J\left(\mathbf{x}^{(k)}\right)$ (например, можно учитывать их симметрию, разреженность и т.п.).

Сравнивая (7.7) с формальным разложением $F(\mathbf{x})$ в ряд Тейлора

$$F(\mathbf{x}) = F\left(\mathbf{x}^{(k)}\right) + F'\left(\mathbf{x}^{(k)}\right)\left(\mathbf{x} - \mathbf{x}^{k}\right) + \frac{1}{2!}F''\left(\mathbf{x}^{(k)}\right)\left(\mathbf{x} - \mathbf{x}^{(k)}\right)^2 + \ldots,$$

видим, что последовательность $\left(\mathbf{x}^{(k)}\right)$ в методе Ньютона получается в результате подмены при каждом $k = 0, 1, 2, \ldots$ нелинейного уравнения $F(\mathbf{x}) = \mathbf{0}$ или, что то же (при достаточной гладкости $F(\mathbf{x})$), уравнения

$$F\left(\mathbf{x}^{(k)}\right) + F'\left(\mathbf{x}^{(k)}\right)\left(\mathbf{x} - \mathbf{x}^{(k)}\right) + \frac{1}{2!}F''\left(\mathbf{x}^{(k)}\right)\left(\mathbf{x} - \mathbf{x}^{(k)}\right)^2 + \ldots = \mathbf{0}$$

линейным уравнением

$$F\left(\mathbf{x}^{(k)}\right) + F'\left(\mathbf{x}^{(k)}\right)\left(\mathbf{x} - \mathbf{x}^{(k)}\right) = \mathbf{0},$$

т.е. пошаговой линеаризацией*). Как следствие этого факта, более тщательно изученного в $\mathbf{R}_1$ (см. § 5.4), можно рассчитывать, что при достаточной гладкости $F(\mathbf{x})$ и достаточно хорошем начальном приближении $\mathbf{x}^{(0)}$ сходимость порождаемой методом Ньютона последовательности $\left(\mathbf{x}^{(k)}\right)$ к решению $\mathbf{x}^*$ будет квадратичной и в многомерном случае. Имеется ряд теорем, устанавливающих это при тех или иных предположениях (см. [13, 61, 80, 129 и др.]). В частности, одна из таких теорем приводится ниже (теорема 7.5 в § 7.8).

Новым, по сравнению со скалярным случаем, фактором, осложняющим применение метода Ньютона к решению n-мерных систем, является необходимость решения n-мерных линейных задач на каждой итерации (обращения матриц в (7.6) или решения СЛАУ в (7.7)), вычислительные затраты на которые растут с ростом n, вообще говоря, непропорционально быстро. Уменьшение таких затрат — одно из направлений модификации метода Ньютона.

Если матрицу Якоби $F'(\mathbf{x})$ вычислить и обратить лишь один раз — в начальной точке $\mathbf{x}^{(0)}$, то от метода Ньютона (7.6)

*) Обратим внимание на некоторую сознательную некорректность использования здесь термина «*линейный*», идущую от «школьной» привычки называть функцию $y = ax + b$ линейной, хотя она не является линейным оператором, ибо не удовлетворяет условию однородности. В [68] в таких случаях используется термин «*аффинная аппроксимация*».

придем к ***модифицированному методу Ньютона****)

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left[F'\left(\mathbf{x}^{(0)}\right)\right]^{-1} F\left(\mathbf{x}^{(k)}\right). \qquad (7.8)$$

Этот метод требует значительно меньших вычислительных затрат на один итерационный шаг, но итераций при этом может потребоваться значительно больше для достижения заданной точности по сравнению с основным методом Ньютона (7.6), поскольку, являясь частным случаем МПИ (с $\mathbf{A} := \left[F'\left(\mathbf{x}^{(0)}\right)\right]^{-1}$), он имеет лишь скорость сходимости геометрической прогрессии**).

Компромиссный вариант — это вычисление и обращение матриц Якоби не на каждом итерационном шаге, а через несколько шагов (иногда такие методы называют ***рекурсивными*** [176]).

Например, простое чередование основного (7.6) и модифицированного (7.8) методов Ньютона приводит к итерационной формуле

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{A}_k F\left(\mathbf{x}^{(k)}\right) - \mathbf{A}_k F\left(\mathbf{x}^{(k)} - \mathbf{A}_k F\left(\mathbf{x}^{(k)}\right)\right), \qquad (7.9)$$

где $\mathbf{A}_k := \left[F'\left(\mathbf{x}^{(k)}\right)\right]^{-1}$, $k = 0, 1, 2, \ldots$. За $\mathbf{x}^{(k)}$ здесь принимается результат последовательного применения одного шага основного, а затем одного шага модифицированного метода, т.е. ***двухступенчатого процесса***

$$\begin{cases} \mathbf{z}^{(k)} = \mathbf{x}^{(k)} - \mathbf{A}_k F\left(\mathbf{x}^{(k)}\right), \\ \mathbf{x}^{(k+1)} = \mathbf{z}^{(k)} - \mathbf{A}_k F\left(\mathbf{z}^{(k)}\right). \end{cases} \qquad (7.10)$$

Доказано, что такой процесс при определенных условиях порождает кубически сходящуюся последовательность $\left(\mathbf{x}^{(k)}\right)$.

Задачу обращения матриц Якоби на каждом k-м шаге метода Ньютона (7.6) можно попытаться решать не точно, а приближенно. Для этого можно применить, например, итерационный процесс Шульца (см. § 3.6), ограничиваясь минимумом — всего одним шагом процесса второго порядка, в котором за начальную матрицу принимается матрица, полученная в результате предыдущего $(k-1)$-го шага. Таким образом приходим к ***методу***

---

*) В книге [181] так называется совсем другая модификация, связанная с оптимизацией шага в «ньютоновском направлении».

**) Независимо от МПИ линейная сходимость модифицированного метода Ньютона (7.8) обосновывается в § 7.8 теоремой 7.6.

Ньютона с последовательной аппроксимацией обратных матриц:

$$\begin{cases} \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{A}_k F\left(\mathbf{x}^{(k)}\right), \\ \mathbf{\Psi}_k = \mathbf{E} - F'\left(\mathbf{x}^{(k+1)}\right)\mathbf{A}_k, \quad \mathbf{A}_{k+1} = \mathbf{A}_k + \mathbf{A}_k \mathbf{\Psi}_k, \end{cases} \qquad (7.11)$$

где $k = 0, 1, 2, \ldots$, а $\mathbf{x}^{(0)}$ и $\mathbf{A}_0$ — начальные вектор и матрица $\left(\approx \left[F'\left(\mathbf{x}^{(0)}\right)\right]^{-1}\right)$ соответственно*). Этот метод (будем называть его более коротко ААМН — ***аппроксимационный аналог метода Ньютона***) имеет простую схему вычислений — поочередное выполнение векторных в первой строке и матричных во второй строке его записи (7.11) операций. Скорость его сходимости почти так же высока, как и у метода Ньютона. Как будет показано в § 7.8, последовательность $\left(\mathbf{x}^{(k)}\right)$ может квадратично сходиться к решению $\mathbf{x}^*$ уравнения $F(\mathbf{x}) = \mathbf{0}$ (при этом матричная последовательность $(\mathbf{A}_k)$ также квадратично сходится к $\mathbf{A}^* := \left[F'\left(\mathbf{x}^*\right)\right]^{-1}$, т.е. в нормально развивающемся итерационном процессе (7.11) должна наблюдаться достаточно быстрая сходимость $(\|\mathbf{\Psi}_k\|)$ к нулю).

Применение той же последовательной аппроксимации обратных матриц к простейшему рекурсивному методу Ньютона (7.9) или, что то же, к двухступенчатому процессу (7.10) определяет его аппроксимационный аналог

$$\begin{cases} \mathbf{z}^{(k)} = \mathbf{x}^{(k)} - \mathbf{A}_k F\left(\mathbf{x}^{(k)}\right), \quad \mathbf{x}^{(k+1)} = \mathbf{z}^{(k)} - \mathbf{A}_k F\left(\mathbf{z}^{(k)}\right), \\ \mathbf{\Psi}_k = \mathbf{E} - F'\left(\mathbf{x}^{(k+1)}\right)\mathbf{A}_k, \quad \mathbf{A}_{k+1} = \mathbf{A}_k + \mathbf{A}_k \mathbf{\Psi}_k, \end{cases} \qquad (7.12)$$

который, как и (7.9), также можно отнести к методам третьего порядка. Доказательство кубической сходимости этого метода требует уже более жестких ограничений на свойства $F(\mathbf{x})$ и близость $\mathbf{x}^{(0)}$ к $\mathbf{x}^*$, $\mathbf{A}_0$ к $\left[F'\left(\mathbf{x}^{(0)}\right)\right]^{-1}$, чем в предыдущем методе. Заметим, что к улучшению сходимости здесь может привести повышение порядка аппроксимации обратных матриц, например, за

---

*) Требования к степени близости $\mathbf{A}_0$ к $\left[F'\left(\mathbf{x}^{(0)}\right)\right]^{-1}$, наряду с другими требованиями, гарантирующими сходимость метода, см. далее в теореме 7.7 и следствии 7.1.

счет добавления еще одного слагаемого в формуле для подсчета $\mathbf{A}_{k+1}$ (см. (3.34)):

$$\mathbf{A}_{k+1} = \mathbf{A}_k + \mathbf{A}_k \mathbf{\Psi}_k + \mathbf{A}_k \mathbf{\Psi}_k^2.$$

На базе метода Ньютона (7.6) можно построить близкий к нему по поведению итерационный процесс, не требующий вычисления производных. Сделаем это, заменив частные производные в матрице Якоби $J(\mathbf{x})$ разностными отношениями, т.е. подставив в формулу (7.5) вместо $\mathbf{A}_k$ матрицу $\left[J\left(\mathbf{x}^{(k)}, \mathbf{h}^{(k)}\right)\right]^{-1}$, где

$$J(\mathbf{x}, \mathbf{h}) := \left(\frac{f_i(x_1, \dots, x_j + h_j, \dots, x_n) - f_i(x_1, \dots, x_j, \dots, x_n)}{h_j} \right)_{i,\, j=1}^{n}.$$

При удачном задании последовательности малых векторов $\mathbf{h}^{(k)} = \left(h_1^{(k)}, \dots, h_n^{(k)}\right)^T$ (постоянной или сходящейся к нулю) полученный таким путем ***разностный*** (или иначе, ***дискретный***) метод Ньютона

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left[J\left(\mathbf{x}^{(k)}, \mathbf{h}^{(k)}\right)\right]^{-1} F\left(\mathbf{x}^{(k)}\right) \qquad (7.13)$$

имеет сверхлинейную, вплоть до квадратичной, скорость сходимости и обобщает на многомерный случай метод (5.29). При задании векторного параметра $\mathbf{h}$ — шага дискретизации — следует учитывать точность машинных вычислений ($macheps$), точность вычисления значений функций f_i, средние значения получаемых приближений (см. [68]).

Можно связать задание последовательности $\left(\mathbf{h}^{(k)}\right)$ с какой-либо сходящейся к нулю векторной последовательностью, например, с последовательностью *невязок* $\left(F\left(\mathbf{x}^{(k)}\right)\right)$ или *поправок* $\left(\mathbf{p}^{(k)}\right)$. Так, полагая $h_j^{(k)} := x_j^{(k-1)} - x_j^{(k)}$, где $j = 1, \dots, n$, а $k = 1, 2, \dots$, приходим к ***простейшему методу секущих*** — обобщению скалярного метода секущих (5.32):

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left[B\left(\mathbf{x}^{(k)}, \mathbf{x}^{(k-1)}\right)\right]^{-1} F\left(\mathbf{x}^{(k)}\right), \qquad (7.14)$$

где

$$B\left(\mathbf{x}^{(k)}, \mathbf{x}^{(k-1)}\right) := \left(\frac{f_i\left(x_1^{(k)}, \dots, x_j^{(k-1)}, \dots, x_n^{(k)}\right) - f_i\left(x_1^{(k)}, \dots, x_j^{(k)}, \dots, x_n^{(k)}\right)}{x_j^{(k-1)} - x_j^{(k)}} \right)_{i,\, j=1}^{n},$$

$k = 1, 2, 3, \dots$.

Этот метод является двухшаговым и требует задания двух начальных точек $\mathbf{x}^{(0)}$ и $\mathbf{x}^{(1)}$. Как было показано в гл.5, при $n = 1$ сходимость метода (7.14) имеет порядок $\frac{1+\sqrt{5}}{2}$. Можно рассчитывать на такую же скорость и в многомерном случае.

К методу секущих так же, как и к методу Ньютона, можно применить пошаговую аппроксимацию обратных матриц на основе метода Шульца. Расчетные формулы этой модификации легко выписать, заменив в совокупности формул ААМН (7.11) матрицу $F'\left(\mathbf{x}^{(k+1)}\right)$ на матрицу $B\left(\mathbf{x}^{(k+1)}, \mathbf{x}^{(k)}\right)$ из (7.14).

Замечание 7.1. Если в одномерном случае разные подходы к линеаризации $f(x)$ приводили к одной и той же формуле секущих (5.32), то для функции n переменных $F(\mathbf{x})$ известно несколько разных обобщений этой формулы в зависимости от того, на какую основу положена линеаризация $F(\mathbf{x})$ в текущей точке. Предложенный здесь простейший метод секущих (7.14) является одним из семейства методов секущих, базирующихся на аппроксимации матриц Якоби. Линейная интерполяция $F(\mathbf{x})$ в $\mathbf{R}_n$ может привести к ряду других методов секущих (см., например, [139]). Среди множества таких методов особый интерес представляет *метод секущих Бройдена*; этому методу посвящен § 7.4.

Замечание 7.2. Для останова процесса вычислений в быстросходящихся методах таких, как метод Ньютона, методы секущих и т.п., часто вполне успешно применяют простой критерий:

$$\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| < \varepsilon \quad \Rightarrow \quad stop, \quad \mathbf{x}^* \approx \mathbf{x}^{(k)}. \qquad (7.15)$$

Это можно объяснить двумя причинами. Во-первых, оценки погрешности здесь довольно «дороги». Имеется в виду как их получение (особенно для различных модификаций базовых методов), так и их реальное применение. Во-вторых, в силу своей быстрой сходимости, к моменту достижения требуемой малости нормы поправки эти методы набирают такую инерцию, что зачастую «проскакивают» установленный порог точности, т.е. выход по критерию (7.15) дает значение $\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\|$ значительно (иногда на несколько порядков) меньшее, чем ε, см. пример в § 7.7.*) Отслеживать факт сходимости в процессе итераций для того, чтобы реагировать на

---

*) Как отмечается в [68], для метода Ньютона установлены неравенства

$$0.5\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\| \le \left\|\mathbf{x}^* - \mathbf{x}^{(k-1)}\right\| \le 2\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right\|,$$

подтверждающие высказанные соображения. В соответствии с этими неравенствами срабатывание критерия (7.15) для метода Ньютона означает, что уже вектор $\mathbf{x}^{(k-1)}$ может быть принят за решение $\mathbf{x}^*$, но поскольку подсчитан вектор $\mathbf{x}^{(k)}$, полагаем

$$\mathbf{x}^* :\approx \mathbf{x}^{(k)}.$$

возможную расходимость в случаях, когда заранее не обеспечены условия сходимости применяемого метода, можно с помощью текущих проверок на уменьшение от шага к шагу поправок и невязок, т.е. выполнение неравенств

$$\left\|\mathbf{x}^{(k)}-\mathbf{x}^{(k-1)}\right\|<\left\|\mathbf{x}^{(k-1)}-\mathbf{x}^{(k-2)}\right\| \quad \text{и} \quad \left\|F\left(\mathbf{x}^{(k)}\right)\right\|<\left\|F\left(\mathbf{x}^{(k-1)}\right)\right\|. \qquad (7.16)$$

7.3. МЕТОД БРАУНА

В отличие от пошаговой линеаризации векторной функции $F(\mathbf{x})$, приведшей к методу Ньютона (7.6), Брауном (1966 г.)*) предложено проводить на каждом итерационном шаге поочередную линеаризацию компонент вектор-функции $F(\mathbf{x})$, т.е. линеаризовать в системе (7.1) сначала функцию f_1, затем f_2 и т.д., и последовательно решать получаемые таким образом уравнения. Чтобы не затенять эту идею громоздкими выкладками и лишними индексами, рассмотрим вывод расчетных формул метода Брауна в двумерном случае.

Пусть требуется найти решение системы

$$\begin{cases} f(x, y)=0, \\ g(x, y)=0, \end{cases} \qquad (7.17)$$

и пусть уже получены приближения x_k, y_k.

Подменим первое уравнение системы (7.17) линейным, полученным по формуле Тейлора для функции двух переменных:

$$f(x, y) \approx f(x_k, y_k)+f'_x(x_k, y_k)(x-x_k)+f'_y(x_k, y_k)(y-y_k)=0.$$

Отсюда выражаем x (обозначим этот результат через $\widetilde{x}$):

$$\widetilde{x}=x_k-\frac{1}{f'_x(x_k, y_k)}\left[f(x_k, y_k)+f'_y(x_k, y_k)(y-y_k)\right]. \qquad (7.18)$$

При $y=y_k$ находим значение $\widetilde{x}_k$ переменной $\widetilde{x}$:

$$\widetilde{x}_k=x_k-\frac{f(x_k, y_k)}{f'_x(x_k, y_k)},$$

которое будем считать лишь промежуточным приближением (т.е. не x_{k+1}), поскольку оно не учитывает второго уравнения системы (7.17).

---

*) Ссылки на первоисточники можно найти в [139].

Подставив в $g(x, y)$ вместо x переменную $\widetilde{x}=\widetilde{x}(y)$, придем к некоторой функции $G(y):=g(\widetilde{x}, y)$ только одной переменной y. Это позволяет линеаризовать второе уравнение системы (7.17) с помощью формулы Тейлора для функции одной переменной:

$$g(\widetilde{x}, y) \approx G(y_k)+G'(y_k)(y-y_k)=0. \qquad (7.19)$$

При нахождении производной $G'(y)$ нужно учесть, что $G(y)=g(\widetilde{x}(y), y)$ есть сложная функция одной переменной y, т.е. применить формулу полной производной

$$G'(y)=g'_x(\widetilde{x}, y)\cdot\widetilde{x}'_y+g'_y(\widetilde{x}, y).$$

Дифференцируя по y равенство (7.18), получаем выражение

$$\widetilde{x}'_y=-\frac{f'_y(x_k, y_k)}{f'_x(x_k, y_k)},$$

подстановка которого в предыдущее равенство при $y=y_k$, $\widetilde{x}=\widetilde{x}_k$ дает

$$G'(y_k)=-g'_k(\widetilde{x}_k, y_k)\cdot\frac{f'_y(x_k, y_k)}{f'_x(x_k, y_k)}+g'_y(\widetilde{x}_k, y_k).$$

При известных значениях $G(y_k)=g(\widetilde{x}_k, y_k)$ и $G'(y_k)$ теперь можно разрешить линейное уравнение (7.19) относительно y (назовем полученное значение y_{k+1}):

$$y_{k+1}=y_k-\frac{G(y_k)}{G'(y_k)}=y_k-$$

$$-\frac{g(\widetilde{x}_k, y_k)f'_x(x_k, y_k)}{f'_x(x_k, y_k)g'_y(\widetilde{x}_k, y_k)-f'_y(x_k, y_k)g'_x(\widetilde{x}_k, y_k)}.$$

Заменяя в (7.18) переменную y найденным значением y_{k+1}, приходим к значению x_{k+1}:

$$x_{k+1}=\widetilde{x}(y_{k+1})=$$

$$=x_k-\frac{1}{f'_x(x_k, y_k)}\left[f(x_k, y_k)+f'_y(x_k, y_k)(y_{k+1}-y_k)\right].$$

Таким образом, реализация метода Брауна решения двумерных нелинейных систем вида (7.17) сводится к следующему.

При выбранных начальных значениях x_0, y_0 каждое последующее приближение по ***методу Брауна*** находится при $k = 0, 1, 2, \ldots$ с помощью совокупности формул

$$\widetilde{x}_k = x_k - \frac{f(x_k, y_k)}{f'_x(x_k, y_k)},$$

$$q_k = \frac{g(\widetilde{x}_k, y_k) \cdot f'_x(x_k, y_k)}{f'_x(x_k, y_k) g'_y(x_k, y_k) - f'_y(x_k, y_k) g'_x(\widetilde{x}_k, y_k)},$$

$$p_k = \frac{f(x_k, y_k) - q_k f'_y(x_k, y_k)}{f'_x(x_k, y_k)},$$

$$x_{k+1} = x_k - p_k, \quad y_{k+1} = y_k - q_k,$$

счет по которым должен выполнятся в той очередности, в которой они записаны.

Вычисления в методе Брауна естественно заканчивать, когда выполнится неравенство $\max\{|p_{k-1}|, |q_{k-1}|\} < \varepsilon$ (с результатом $(x^*, y^*) :\approx (x_k, y_k)$). В ходе вычислений следует контролировать немалость знаменателей расчетных формул. Заметим, что функции f и g в этом методе неравноправны, и перемена их ролями может изменить ситуацию со сходимостью.

Указывая на наличие *квадратичной сходимости* метода Брауна, в [139] отмечают, что рассчитывать на его бóльшую по сравнению с методом Ньютона эффективность в смысле вычислительных затрат можно лишь в случае, когда фигурирующие в нем частные производные заменяются разностными отношениями.

7.4. МЕТОД СЕКУЩИХ БРОЙДЕНА

Чтобы приблизиться к пониманию идей, лежащих в основе предлагаемого вниманию метода, вернемся сначала к изучавшемуся в двух предыдущих главах одномерному случаю.

В процессе построения методов Ньютона и секущих решения нелинейного скалярного уравнения.

$$f(x) = 0 \tag{7.20}$$

функция $f(x)$ в окрестности текущей точки x_k подменяется линейной функцией (***аффинной моделью***)

$$\varphi_k(a_k, x) := f(x_k) + a_k(x - x_k). \tag{7.21}$$

Приравнивание к нулю последней, т.е. решение линейного

уравнения

$$f(x_k) + a_k(x - x_k) = 0,$$

порождает итерационную формулу

$$x_{k+1} = x_k - a_k^{-1} f(x_k) \tag{7.22}$$

для вычисления приближений к корню уравнения (7.20).

Если потребовать, чтобы заменяющая функцию $f(x)$ вблизи точки x_k аффинная модель $\varphi_k(a_k, x)$ имела в этой точке одинаковую с ней производную, то, дифференцируя (7.21), получаем значение коэффициента

$$a_k = f'(x_k),$$

подстановка которого в (7.22) приводит к известному методу Ньютона (5.14). Если же исходить из того, что наряду с равенством $\varphi_k(a_k, x_k) = f(x_k)$ должно иметь место совпадение функций $f(x)$ и $\varphi_k(a_k, x)$ в предшествующей x_k точке x_{k-1}, т.е. из равенства

$$\varphi_k(a_k, x_{k-1}) = f(x_{k-1}),$$

или, в соответствии с (7.21),

$$f(x_k) + a_k(x_{k-1} - x_k) = f(x_{k-1}), \tag{7.23}$$

то получаем коэффициент

$$a_k = \frac{f(x_{k-1}) - f(x_k)}{x_{k-1} - x_k},$$

превращающий (7.22) в известную формулу секущих (5.32).

Равенство (7.23), переписанное в виде

$$a_k(x_{k-1} - x_k) = f(x_{k-1}) - f(x_k),$$

называют ***соотношением секущих в*** $\mathbf{R}_1$ [68]. Оно легко обобщается на n-мерный случай и лежит в основе вывода метода Бройдена. Опишем этот вывод.

В n-мерном векторном пространстве $\mathbf{R}_n$ соотношение секущих представляется равенством

$$\mathbf{B}_k(\mathbf{x}^{(k-1)} - \mathbf{x}^{(k)}) = F(\mathbf{x}^{(k-1)}) - F(\mathbf{x}^{(k)}), \tag{7.24}$$

где $\mathbf{x}^{(k-1)}$, $\mathbf{x}^{(k)}$ — известные n-мерные векторы, $F: \mathbf{R}_n \to \mathbf{R}_n$ — данное нелинейное отображение, а $\mathbf{B}_k$ — некоторая матрица линейного преобразования в $\mathbf{R}_n$. С обозначениями

$$\mathbf{s}^{(k)} := \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}, \quad \mathbf{y}^{(k)} := F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k-1)}), \tag{7.25}$$

соотношение секущих в $\mathbf{R}_n$ обретает более короткую запись:

$$\mathbf{B}_k \mathbf{s}^{(k)} = \mathbf{y}^{(k)}. \tag{7.24a}$$

Аналогично одномерному случаю, а именно, по аналогии с формулой (7.22), будем искать приближения к решению $\mathbf{x}^*$ векторного уравнения (7.1а) по формуле

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{B}_k^{-1} F(\mathbf{x}^{(k)}). \tag{7.26}$$

Желая, чтобы эта формула обобщала метод секущих (5.32), обратимую $n \times n$-матрицу $\mathbf{B}_k$ в ней нужно подобрать так, чтобы она удовлетворяла соотношению секущих (7.24). Но это соотношение не определяет однозначно матрицу $\mathbf{B}_k$: глядя на равенство (7.24a), легко понять, что при $n > 1$ существует множество матриц $\mathbf{B}_k$, преобразующих заданный n-мерный вектор $\mathbf{s}^{(k)}$ в другой заданный вектор $\mathbf{y}^{(k)}$ (отсюда — ясность в понимании того, что могут быть различные обобщения одномерного метода секущих, см. в замечание 7.1).

При формировании матрицы $\mathbf{B}_k$ будем рассуждать следующим образом.

Переходя от имеющейся в точке $\mathbf{x}^{(k-1)}$ аффинной модели функции $F(\mathbf{x})$

$$\mathbf{\Phi}_{k-1} := F\left(\mathbf{x}^{(k-1)}\right) + \mathbf{B}_{k-1}\left(\mathbf{x} - \mathbf{x}^{(k-1)}\right) \tag{7.27}$$

к такой же модели в точке $\mathbf{x}^{(k)}$

$$\mathbf{\Phi}_k = F\left(\mathbf{x}^{(k)}\right) + \mathbf{B}_k\left(\mathbf{x} - \mathbf{x}^{(k)}\right), \tag{7.28}$$

мы не имеем о матрице линейного преобразования $\mathbf{B}_k$ никаких сведений, кроме соотношения секущих (7.24). Поэтому исходим из того, что при этом переходе изменения в модели должны быть минимальными. Эти изменения характеризуют разность $\mathbf{\Phi}_k - \mathbf{\Phi}_{k-1}$. Вычтем из равенства (7.28) определяющее $\mathbf{\Phi}_{k-1}$ равенство (7.27) и преобразуем результат, привлекая соотношение секущих (7.24). Имеем:

$$\mathbf{\Phi}_k - \mathbf{\Phi}_{k-1} = F\left(\mathbf{x}^{(k)}\right) - F\left(\mathbf{x}^{(k-1)}\right) + \mathbf{B}_k(\mathbf{x} - \mathbf{x}^{(k)}) - \mathbf{B}_{k-1}(\mathbf{x} - \mathbf{x}^{(k-1)}) =$$

$$= \mathbf{B}_k\left(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\right) - \mathbf{B}_k \mathbf{x}^{(k)} + \mathbf{B}_{k-1}\mathbf{x}^{(k-1)} + (\mathbf{B}_k - \mathbf{B}_{k-1})\mathbf{x} =$$

$$= (\mathbf{B}_k - \mathbf{B}_{k-1})\left(\mathbf{x} - \mathbf{x}^{(k-1)}\right).$$

Представим вектор $\mathbf{x} - \mathbf{x}^{(k-1)}$ в виде линейной комбинации фик-

сированного вектора $\mathbf{s}^{(k)}$, определенного в (7.25), и некоторого вектора $\mathbf{t}$, ему ортогонального:

$$\mathbf{x} - \mathbf{x}^{(k-1)} = \alpha \mathbf{s}^{(k)} + \mathbf{t}, \quad \alpha \in \mathbf{R}_1, \ \ \mathbf{t} \in \mathbf{R}_n: \ \left(\mathbf{t}, \mathbf{s}^{(k)}\right) = 0.$$

Подстановкой этого представления вектора $\mathbf{x} - \mathbf{x}^{(k-1)}$ в разность $\mathbf{\Phi}_k - \mathbf{\Phi}_{k-1}$ получаем другой ее вид

$$\mathbf{\Phi}_k - \mathbf{\Phi}_{k-1} = \alpha(\mathbf{B}_k - \mathbf{B}_{k-1})\mathbf{s}^{(k)} + (\mathbf{B}_k - \mathbf{B}_{k-1})\mathbf{t}. \tag{7.29}$$

Анализируя выражение (7.29), замечаем, что первое слагаемое в нем не может быть изменено, поскольку

$$(\mathbf{B}_k - \mathbf{B}_{k-1})\mathbf{s}^{(k)} = \mathbf{B}_k \mathbf{s}^{(k)} - \mathbf{B}_{k-1}\mathbf{s}^{(k)} = \mathbf{y}^{(k)} - \mathbf{B}_{k-1}\mathbf{s}^{(k)}$$

— фиксированный вектор при фиксированном k. Поэтому минимальному изменению аффинной модели $\mathbf{\Phi}_{k-1}$ будет отвечать случай, когда второе слагаемое в (7.29) будет нуль-вектором при всяких векторах $\mathbf{t}$, ортогональных векторам $\mathbf{s}^{(k)}$, т.е. $\mathbf{B}_k$ следует находить из условия

$$(\mathbf{B}_k - \mathbf{B}_{k-1})\mathbf{t} = 0 \quad \forall \mathbf{t} : \left(\mathbf{t}, \mathbf{s}^{(k)}\right) = 0. \tag{7.30}$$

Непосредственной проверкой убеждаемся, что условие (7.30) будет выполнено, если матричную поправку $\mathbf{B}_k - \mathbf{B}_{k-1}$ взять в виде одноранговой $n \times n$-матрицы

$$\mathbf{B}_k - \mathbf{B}_{k-1} = \frac{\left(\mathbf{y}^{(k)} - \mathbf{B}_{k-1}\mathbf{s}^{(k)}\right)\left(\mathbf{s}^{(k)}\right)^T}{\left(\mathbf{s}^{(k)}\right)^T \mathbf{s}^{(k)}}.$$

Таким образом, приходим к так называемой ***формуле пересчета С. Бройдена*** (1965 г.)

$$\mathbf{B}_k = \mathbf{B}_{k-1} + \frac{\left(\mathbf{y}^{(k)} - \mathbf{B}_{k-1}\mathbf{s}^{(k)}\right)\left(\mathbf{s}^{(k)}\right)^T}{\left(\mathbf{s}^{(k)}\right)^T \mathbf{s}^{(k)}}, \tag{7.31}$$

которая позволяет простыми вычислениями перейти от старой матрицы $\mathbf{B}_{k-1}$ к новой $\mathbf{B}_k$ такой, чтобы выполнялось соотношение секущих (7.24a) в новой точке и при этом изменения в аффинной модели (7.27) были минимальны (строго доказано [68], что такое построение отвечает минимальности поправки $\mathbf{B}_k - \mathbf{B}_{k-1}$ по норме Фробениуса на множестве матриц $\mathbf{B}_k$, удовлетворяющих соотношению секущих (7.24a)).

Совокупность формул (7.26), (7.31) вместе с обозначениями (7.25) называют ***методом секущих Бройдена*** или просто

методом Бройдена решения систем нелинейных числовых уравнений*).

Хотя в методах секущих обычным является задание двух начальных векторов ($\mathbf{x}^{(0)}$ и $\mathbf{x}^{(1)}$), для метода Бройдена характерно другое начало итерационного процесса. Здесь нужно задать один начальный вектор $\mathbf{x}^{(0)}$, начальную матрицу $\mathbf{B}_0$ и далее в цикле по $k = 0, 1, 2, \ldots$ последовательно выполнять следующие операции:

1. *решить линейную систему*

$$\mathbf{B}_k \mathbf{s}^{(k+1)} = -F\left(\mathbf{x}^{(k)}\right) \tag{7.32}$$

относительно вектора $\mathbf{s}^{(k+1)}$ (см. (7.26));

2. *найти векторы* $\mathbf{x}^{(k+1)}$ *и* $\mathbf{y}^{(k+1)}$ (см. (7.25)):

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{s}^{(k+1)}, \quad \mathbf{y}^{(k+1)} = F\left(\mathbf{x}^{(k+1)}\right) - F\left(\mathbf{x}^{(k)}\right); \tag{7.33}$$

3. *сделать проверку на останов* (например, с помощью проверки на малость величин $\left\|\mathbf{s}^{(k+1)}\right\|$ и/или $\left\|\mathbf{y}^{(k+1)}\right\|$) *и, если нужная точность не достигнута, вычислить новую матрицу* $\mathbf{B}_k$ *по формуле пересчета* (см. (7.31))

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \frac{\left(\mathbf{y}^{(k+1)} - \mathbf{B}_k \mathbf{s}^{(k+1)}\right)\left(\mathbf{s}^{(k+1)}\right)^T}{\left(\mathbf{s}^{(k+1)}\right)^T \mathbf{s}^{(k+1)}}. \tag{7.34}$$

В качестве матрицы $\mathbf{B}_0$, требуемой равенством (7.32) для запуска итерационного процесса Бройдена, чаще всего берут матрицу Якоби $F'(\mathbf{x}^{(0)})$ или какую-нибудь ее аппроксимацию. При этом, как отмечается в [68], получаемые далее пересчетом (7.34) матрицы $\mathbf{B}_1$, $\mathbf{B}_2$, ... не всегда можно считать близкими к соответствующим матрицам Якоби $F'\left(\mathbf{x}^{(1)}\right)$, $F'\left(\mathbf{x}^{(2)}\right)$, ... (что может иногда сыграть полезную роль при вырождении матриц $F'\left(\mathbf{x}^*\right)$). Но, в то же время, показывается, что при определенных требованиях к матрицам Якоби $F'(\mathbf{x})$ матрицы $\mathbf{B}_k$ обладают «свойством ограниченного ухудшения», означающим, что если и

---

*) Имеются также различные варианты метода Бройдена, успешно применяемые к задачам безусловной оптимизации.

происходит увеличение $\left\|\mathbf{B}_k - F'\left(\mathbf{x}^{(k)}\right)\right\|$ с увеличением номера итерации k, то достаточно медленно. С помощью этого свойства доказываются утверждения о *линейной сходимости* $\left(\mathbf{x}^{(k)}\right)$ к $\mathbf{x}^*$ при достаточной близости $\mathbf{x}^{(0)}$ к $\mathbf{x}^*$ и $\mathbf{B}_0$ к $F'\left(\mathbf{x}^{(0)}\right)$, а в тех предположениях, при которых можно доказать квадратичную сходимость метода Ньютона (7.6), — о *сверхлинейной сходимости последовательности приближений по методу Бройдена.*

Как и в случаях применения других методов решения нелинейных систем, проверка выполнимости каких-то условий сходимости итерационного процесса Бройдена весьма затруднительна и обычно заменяется проверками на выполнимость неравенств типа (7.16).

Формуле пересчета (7.34) в итерационном процессе Бройдена можно придать чуть более простой вид.

Так как, в силу (7.32) и (7.33),

$$\mathbf{y}^{(k+1)} - \mathbf{B}_k \mathbf{s}^{(k+1)} = F\left(\mathbf{x}^{(k+1)}\right) - F\left(\mathbf{x}^{(k)}\right) + F\left(\mathbf{x}^{(k)}\right) = F\left(\mathbf{x}^{(k+1)}\right),$$

а

$$\left(\mathbf{s}^{(k+1)}\right)^T \mathbf{s}^{(k+1)} = \left(\mathbf{s}^{(k+1)}, \mathbf{s}^{(k+1)}\right) = \left\|\mathbf{s}^{(k+1)}\right\|_2^2 = \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|_2^2,$$

то из формулы (7.34) получаем формально эквивалентную ей формулу пересчета

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \frac{F\left(\mathbf{x}^{(k+1)}\right)\left(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right)^T}{\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|_2^2}, \tag{7.35}$$

которую можно использовать вместо (7.34) в совокупности с формулой (7.26) или с (7.32), (7.33) (без вычисления вектора $\mathbf{y}^{(k+1)}$). Такое преобразование итерационного процесса Бройдена несколько сокращает объем вычислений (на одно матрично-векторное умножение на каждой итерации). Не следует, правда, забывать, что при замене формулы (7.34) формулой (7.35) может измениться ситуация с вычислительной устойчивостью метода; к счастью, это случается здесь крайне редко, а именно, в тех случаях, когда для получения решения с нужной точностью требуется много итераций по методу Бройдена, т.е. когда и применять его не стоит.

7.5. ОБОБЩЕНИЕ ПОЛЮСНОГО МЕТОДА НЬЮТОНА НА МНОГОМЕРНЫЙ СЛУЧАЙ

В предыдущем параграфе при выводе метода секущих Бройдена решения систем нелинейных уравнений вида (7.1) мы возвращались к одномерной задаче (7.20) и рассматривали интерпретацию метода секущих, отличную от данной ранее в § 5.6. Поступим аналогично и здесь, переложив вывод одномерного полюсного метода Ньютона (5.36) на векторную основу. Будем при этом руководствоваться теми же, что и в § 5.7, геометрическими соображениями, опирающимися на рис. 5.12, и пользоваться теми же обозначениями.

Касательную к кривой $y = f(x)$ в точке $(x_k; f(x_k))$ зададим условием ортогональности текущего вектора*) $\mathbf{u} := (x - x_k; y - f(x_k))$ этой прямой и ее нормального вектора, в качестве которого можно взять вектор $\mathbf{n} := (f'(x_k); -1)$. Уравнение прямой, проходящей через полюс $P(c; d)$ и связанную с уже известным приближением x_k точку $(x_k; 0)$, получим из условия коллинеарности текущего вектора этой прямой $\mathbf{v} := (x - x_k; y)$ и ее направляющего вектора $\mathbf{l} := (c - x_k; d)$. Таким образом, точку пересечения двух прямых, проекцию которой на ось абсцисс считаем новым приближением x_{k+1}, находим из совокупности условий

$$\mathbf{u} \perp \mathbf{n}, \quad \mathbf{v} \parallel \mathbf{l}. \tag{7.36}$$

Первое из этих условий означает равенство нулю скалярного произведения $(\mathbf{n}, \mathbf{u})$, второе — пропорциональность соответствующих координат векторов $\mathbf{v}$ и $\mathbf{l}$ или, иначе, равенство нулю составленного из них определителя. Следовательно, искомое приближение x_{k+1} есть первая компонента вектора, служащего решением линейной системы

$$\begin{cases} f'(x_k)(x - x_k) - y + f(x_k) = 0, \\ \begin{vmatrix} x - x_k & y \\ c - x_k & d \end{vmatrix} = 0 \end{cases} \tag{7.37}$$

(вторая компонента — ордината точки пересечения указанных прямых — после вычисления значения $f(x_{k+1})$ может дать информацию об отклонении от функции $f(x)$ в точке x_{k+1} ее локальной аффинной модели, каковой является проведенная в точке x_k касательная). Ясно, что получаемое из системы (7.37) значение $x := x_{k+1}$ тождественно его выражению по формуле (5.36).

---

*) Не будем здесь различать двумерные векторы-строки и векторы-столбцы.

Рассмотренный векторный подход к построению одномерного полюсного метода Ньютона служит ключом для его распространения на двумерный случай на основе таких же геометрических, но уже пространственных соображений.

Пусть требуется найти приближенное решение двумерной нелинейной системы (7.17) в предположении непрерывной дифференцируемости входящих в нее функций $f(x, y)$ и $g(x, y)$ в некоторой области G, содержащей искомое решение $\mathbf{x}^* = (x^*; y^*)$ и приближения к нему $\mathbf{x}^{(k)} = (x^{(k)}; y^{(k)})$, $k = 0, 1, 2, \ldots$.

Будем считать, что уже найдено k-е приближение к решению $\mathbf{x}^*$ и нужно получить правило перехода к $(k+1)$-му приближению. В сделанном предположении о гладкости функций $f(x, y)$ и $g(x, y)$ можно провести касательные плоскости в точке $(x_k; y_k; 0)$ к определяемым ими поверхностям

$$z = f(x, y) \quad \text{и} \quad z = g(x, y). \tag{7.38}$$

Эти плоскости задаются текущими векторами

$$\mathbf{v}_1 := \left(x - x^{(k)}; y - y^{(k)}; z - f\left(x^{(k)}, y^{(k)}\right)\right),$$
$$\mathbf{v}_2 := \left(x - x^{(k)}; y - y^{(k)}; z - g\left(x^{(k)}, y^{(k)}\right)\right)$$

и нормалями

$$\mathbf{n}_1 := \left(f'_x\left(x^{(k)}, y^{(k)}\right); f'_y\left(x^{(k)}, y^{(k)}\right); -1\right),$$
$$\mathbf{n}_2 := \left(g'_x\left(x^{(k)}, y^{(k)}\right); g'_y\left(x^{(k)}, y^{(k)}\right); -1\right)$$

соответственно, т.е. аналогично первому из условий (7.36) должно быть $\mathbf{n}_1 \perp \mathbf{v}_1$, $\mathbf{n}_2 \perp \mathbf{v}_2$, иначе,

$$(\mathbf{n}_1, \mathbf{v}_1) = 0, \quad (\mathbf{n}_2, \mathbf{v}_2) = 0. \tag{7.39}$$

Пересечение двух касательных плоскостей, т.е. образ, определяемый уравнениями (7.39), есть прямая в трехмерном пространстве, общая точка которой с координатной плоскостью Oxy является ньютоновским приближением $\tilde{\mathbf{x}}^{(k+1)}$ к решению $\mathbf{x}^*$ сис-

темы (7.17). Наша цель — построить третью плоскость, пересечение которой с упомянутой прямой (линией пересечения касательных плоскостей) давало бы точку в пространстве $\mathbf{R}_3$ такую, проекция которой на плоскость Oxy могла бы оказаться ближе к $\mathbf{x}^*$, чем $\widetilde{\mathbf{x}}^{(k+1)}$.

Чтобы осуществить поставленную цель, зафиксируем в $\mathbf{R}_3$ две несовпадающие между собой и с $K\left(x^{(k)}; y^{(k)}; 0\right)$ точки — ***полюсы*** $P_1(a_1; b_1; d_1)$ и $P_2(a_2; b_2; d_2)$. Через указанные три точки K, P_1, P_2 можно провести единственную плоскость (которая здесь играет роль прямой, проходящей через полюс и точку $(x_k; 0)$ в одномерной ситуации). Взяв текущую точку $M(x; y; z)$ и образовав текущий вектор $\mathbf{v}_3 := \overrightarrow{KM}$ этой третьей плоскости, можно задать ее условием компланарности трех векторов: $\overrightarrow{KP_1}$, $\overrightarrow{KP_2}$ и $\overrightarrow{KM}$ (что служит аналогом второго из условий (7.36)).

Запишем совокупность всех трех описанных средствами векторной алгебры плоскостей в координатной форме. Имеем:

$$\begin{cases} f'_x\left(x^{(k)}, y^{(k)}\right)\left(x - x^{(k)}\right) + f'_y\left(x^{(k)}, y^{(k)}\right)\left(y - y^{(k)}\right) - z + f\left(x^{(k)}, y^{(k)}\right) = 0, \\ g'_x\left(x^{(k)}, y^{(k)}\right)\left(x - x^{(k)}\right) + g'_y\left(x^{(k)}, y^{(k)}\right)\left(y - y^{(k)}\right) - z + g\left(x^{(k)}, y^{(k)}\right) = 0, \\ \begin{vmatrix} x - x^{(k)} & y - y^{(k)} & z \\ a_1 - x^{(k)} & b_1 - y^{(k)} & d_1 \\ a_2 - x^{(k)} & b_2 - y^{(k)} & d_2 \end{vmatrix} = 0. \end{cases}$$

Первые две координаты вектора $(x; y; z)$, служащего решением полученной системы уравнений, считаем искомым приближением $\left(x^{(k+1)}; y^{(k+1)}\right)$. Введя ***поправки***

$$p^{(k)} = x^{(k+1)} - x^{(k)}, \quad q^{(k)} = y^{(k+1)} - y^{(k)}, \tag{7.40}$$

эту систему превращаем в систему уравнений относительно

неизвестных $p^{(k)}$, $q^{(k)}$ и z:

$$\begin{cases} f'_x\left(x^{(k)}, y^{(k)}\right) p^{(k)} + f'_y\left(x^{(k)}, y^{(k)}\right) q^{(k)} - z = -f\left(x^{(k)}, y^{(k)}\right), \\ g'_x\left(x^{(k)}, y^{(k)}\right) p^{(k)} + g'_y\left(x^{(k)}, y^{(k)}\right) q^{(k)} - z = -g\left(x^{(k)}, y^{(k)}\right), \\ \begin{vmatrix} p^{(k)} & q^{(k)} & z \\ a_1 - x^{(k)} & b_1 - y^{(k)} & d_1 \\ a_2 - x^{(k)} & b_2 - y^{(k)} & d_2 \end{vmatrix} = 0. \end{cases} \tag{7.41}$$

Для исключения вспомогательной переменной z из линейной системы (7.41) выразим ее из третьего уравнения. Обозначив

$$\Delta_p^{(k)} := \begin{vmatrix} b_1 - y^{(k)} & d_1 \\ b_2 - y^{(k)} & d_2 \end{vmatrix}, \quad \Delta_q^{(k)} := -\begin{vmatrix} a_1 - x^{(k)} & d_1 \\ a_2 - x^{(k)} & d_2 \end{vmatrix},$$

$$\Delta_z^{(k)} := \begin{vmatrix} a_1 - x^{(k)} & b_1 - y^{(k)} \\ a_2 - x^{(k)} & b_2 - y^{(k)} \end{vmatrix}, \tag{7.42}$$

раскрываем фигурирующий в (7.41) определитель по элементам первой строки:

$$\Delta_p^{(k)} p^{(k)} + \Delta_q^{(k)} q^{(k)} + \Delta_z^{(k)} z = 0.$$

Отсюда находим выражение

$$-z = \frac{\Delta_p^{(k)}}{\Delta_z^{(k)}} p^{(k)} + \frac{\Delta_q^{(k)}}{\Delta_z^{(k)}} q^{(k)}, \tag{7.43}$$

подставляя которое в первые два уравнения системы (7.41), приходим к двумерной линейной системе

$$\begin{cases} \left(f'_x\left(x^{(k)}, y^{(k)}\right) + \dfrac{\Delta_p^{(k)}}{\Delta_z^{(k)}}\right) p^{(k)} + \left(f'_y\left(x^{(k)}, y^{(k)}\right) + \dfrac{\Delta_q^{(k)}}{\Delta_z^{(k)}}\right) q^{(k)} = \\ \qquad = -f\left(x^{(k)}, y^{(k)}\right), \\ \left(g'_x\left(x^{(k)}, y^{(k)}\right) + \dfrac{\Delta_p^{(k)}}{\Delta_z^{(k)}}\right) p^{(k)} + \left(g'_y\left(x^{(k)}, y^{(k)}\right) + \dfrac{\Delta_q^{(k)}}{\Delta_z^{(k)}}\right) q^{(k)} = \\ \qquad = -g\left(x^{(k)}, y^{(k)}\right). \end{cases} \tag{7.44}$$

Фактически эта система вместе с обозначениями (7.42) и определяет ***двумерный полюсный метод Ньютона*** для нелинейной системы (7.17). Найдя из нее поправки $p^{(k)}$, $q^{(k)}$, в соответствии с равенствами (7.40) получаем очередное приближение $\mathbf{x}^{(k+1)}$:

$$x^{(k+1)} = x^{(k)} + p^{(k)}, \quad y^{(k+1)} = y^{(k)} + q^{(k)}.$$

Дальнейшее обобщение полюсного метода Ньютона, т.е. переход от размерности 2 к произвольной размерности $n \geq 2$, совершаем формально на основе предыдущего построения.

Пусть задана нелинейная система (7.1), функции $f_i(\mathbf{x})$ в которой считаем достаточно гладкими. Совокупность всех «касательных гиперплоскостей» к гиперповерхностям, определяемым данными функциями $f_i(\mathbf{x})$ (образующими вектор $F(\mathbf{x}) \in \mathbf{R}_n$) в точке $\mathbf{x}^{(k)} \in \mathbf{R}_n$, можно описать векторно–матричным уравнением

$$F'\left(\mathbf{x}^{(k)}\right)\left(\mathbf{x} - \mathbf{x}^{(k)}\right) - \mathbf{z} = -F\left(\mathbf{x}^{(k)}\right), \tag{7.45}$$

где $\mathbf{z} := (x_{n+1};\ x_{n+1};\ \ldots;\ x_{n+1})^T$ — n-мерный вектор, каждой компонентой которого служит вспомогательная переменная x_{n+1}, входящая в уравнения гиперповерхностей $x_{n+1} = f_i(\mathbf{x})$ (ср. с (7.38)).

Зададим n ***полюсов*** $P_i(c_{i1};\ c_{i2};\ \ldots;\ c_{in};\ d_i)$ $(i = 1, 2, \ldots, n)$ так, чтобы они не принадлежали одной гиперплоскости пространства $\mathbf{R}_{n+1}$. Через все эти полюсы P_i и точку $\left(x_1^{(k)};\ x_2^{(k)};\ \ldots x_n^{(k)};\ 0\right)$, определяемую известным приближением $\mathbf{x}^{(k)}$ к решению системы (7.1), проводим гиперплоскость, уравнение которой аналогично двумерному случаю задаем условием равенства нулю определителя $(n+1)$-го порядка:

$$\begin{vmatrix} x_1 - x_1^{(k)} & x_2 - x_2^{(k)} & \cdots & x_n - x_n^{(k)} & x_{n+1} \\ c_{11} - x_1^{(k)} & c_{12} - x_2^{(k)} & \cdots & c_{1n} - x_n^{(k)} & d_1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ c_{n1} - x_1^{(k)} & c_{n2} - x_2^{(k)} & \cdots & c_{nn} - x_n^{(k)} & d_n \end{vmatrix} = 0. \tag{7.46}$$

Векторно-матричное уравнение (7.45) и скалярное уравнение (7.46), в принципе, уже определяют n-полюсный метод Ньютона для построения приближений к решению системы (7.1). Чтобы

записать соответствующую линейную систему относительно ***поправок***

$$\mathbf{p}^{(k)} = \begin{pmatrix} p_1^{(k)} \\ \cdots \\ p_n^{(k)} \end{pmatrix} := \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} = \begin{pmatrix} x_1^{(k+1)} - x_1^{(k)} \\ \cdots \\ x_n^{(k+1)} - x_n^{(k)} \end{pmatrix} \tag{7.47}$$

(аналогичную системе (7.44) двумерного случая), введем следующие обозначения. Положим

$$\mathbf{C} := \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \cdots & \cdots & \cdots \\ c_{n1} & \cdots & c_{nn} \end{pmatrix}, \quad \mathbf{d} := \begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix},$$

$$\mathbf{X}_k := \begin{pmatrix} \left(\mathbf{x}^{(k)}\right)^T \\ \vdots \\ \left(\mathbf{x}^{(k)}\right)^T \end{pmatrix} = \begin{pmatrix} x_1^{(k)} & \cdots & x_n^{(k)} \\ \cdots & \cdots & \cdots \\ x_1^{(k)} & \cdots & x_n^{(k)} \end{pmatrix}$$

и образуем квадратную $(n+1)$-мерную матрицу следующей структуры:

$$\mathbf{Q}_k := \begin{pmatrix} p_1^{(k)} \cdots p_n^{(k)} & x_{n+1} \\ \mathbf{C} - \mathbf{X}_k & \mathbf{d} \end{pmatrix}.$$

Тогда на основе (7.45), (7.46) имеем $(n+1)$-мерную систему уравнений относительно неизвестных $p_1^{(k)}, \ldots, p_n^{(k)}, x_{n+1}$:

$$\begin{cases} F'\left(\mathbf{x}^{(k)}\right)\mathbf{p}^{(k)} - \mathbf{z} = -F\left(\mathbf{x}^{(k)}\right), \\ \det \mathbf{Q}_k = 0. \end{cases} \tag{7.48}$$

Как и в двумерном случае, из второго уравнения этой системы выражаем вспомогательную неизвестную x_{n+1}:

$$x_{n+1} = -\frac{1}{\Delta^{(k)}} \sum_{j=1}^{n} a_j^{(k)} p_j^{(k)}, \tag{7.49}$$

где $\Delta^{(k)} := (-1)^n \det(\mathbf{C} - \mathbf{X}_k)$, а $a_j^{(k)}$ есть алгебраические дополнения к элементам $p_j^{(k)}$ первой строки матрицы $\mathbf{Q}_k$ (что через соответствующие миноры M_{1j} этой матрицы можно представить так:

$$a_j^{(k)} = (-1)^{1+j} M_{1j}, \quad j = 1, 2, \ldots, n).$$

Заменив в (7.48) все компоненты вектора $\mathbf{z}$ найденным их значением (7.49), приходим к следующему линейному векторно-матричному уравнению относительно вектора-поправки $\mathbf{p}^{(k)}$:

$$\left[F'\left(\mathbf{x}^{(k)}\right)+\mathbf{A}_k\right]\mathbf{p}^{(k)}=-F\left(\mathbf{x}^{(k)}\right), \qquad (7.50)$$

где

$$\mathbf{A}_k:=\frac{(-1)^n}{\det(\mathbf{C}-\mathbf{X}_k)}\begin{pmatrix} a_1^{(k)} & \cdots & a_n^{(k)} \\ \cdots & \cdots & \cdots \\ a_1^{(k)} & \cdots & a_n^{(k)} \end{pmatrix}. \qquad (7.51)$$

Уравнение (7.50) вместе со связью (7.47), согласно которой

$$\mathbf{x}^{(k+1)}=\mathbf{x}^{(k)}+\mathbf{p}^{(k)}, \qquad (7.52)$$

является неявной формой n-полюсного метода Ньютона для уравнения (7.1а).

Совокупности формул (7.50)–(7.52) можно придать другой вид:

$$\mathbf{x}^{(k+1)}=\mathbf{x}^{(k)}-\left[F'\left(\mathbf{x}^{(k)}\right)+\mathbf{A}_k\right]^{-1}F\left(\mathbf{x}^{(k)}\right), \qquad (7.53)$$

который удобно трактовать как явный метод Ньютона со своеобразной коррекцией матриц Якоби путем прибавления к ним формирующихся по заданному правилу матриц $\mathbf{A}_k$. Как и в одномерном случае, для ускорения сходимости последовательности приближений $\mathbf{x}^{(k)}$ полюсы P_i целесообразно изменять в такт с изменением значений функций, и в самом простом случае есть смысл фиксировать матрицу $\mathbf{C}$, а вектор $\mathbf{d}$ $(=\mathbf{d}^{(k)})$ брать равным $F\left(\mathbf{x}^{(k)}\right)$ или $-F\left(\mathbf{x}^{(k)}\right)$.

Замечание 7.3. Обобщение полюсного метода секущих (5.37) на многомерный случай, как и обычного метода секущих (5.32) (о чем уже не единожды упоминалось), может иметь много вариаций. Одним из плодотворных подходов к такому обобщению является использование здесь рассмотренных в предыдущем параграфе формул пересчета Бройдена.

7.6. О РЕШЕНИИ НЕЛИНЕЙНЫХ СИСТЕМ МЕТОДАМИ СПУСКА

Общий недостаток всех рассмотренных выше методов решения систем нелинейных уравнений — это сугубо локальный характер сходимости, затрудняющий их применение в случаях (довольно типичных), когда имеются проблемы с выбором хо-

роших начальных приближений. Помощь здесь может прийти со стороны численных методов оптимизации — ветви вычислительной математики, обычно выделяемой в самостоятельную дисциплину. Для этого нужно поставить задачу нахождения решений данной нелинейной системы как оптимизационную или, иначе, экстремальную задачу. Ради геометрической интерпретации проводимых ниже рассуждений и их результатов, ограничимся, как и в § 7.3, рассмотрением системы, состоящей из двух уравнений с двумя неизвестными, т.е. системы (7.17).

Из функций f и g системы (7.17) образуем новую функцию

$$\Phi(x,y):=f^2(x,y)+g^2(x,y). \qquad (7.54)$$

Так как эта функция неотрицательна, то найдется точка*) $(x^*;\,y^*)$ такая, что

$$\Phi(x,y)\ge\Phi(x^*,y^*)\ge 0 \quad \forall\,(x;y)\in\mathbf{R}_2,$$

т.е. $(x^*;\,y^*)=\arg\min\limits_{x\in\mathbf{R}_2}\Phi(x,y)$. Следовательно, если тем или иным способом удается получить точку $(x^*;\,y^*)$, минимизирующую функцию $\Phi(x,y)$, и если при этом окажется, что $\min\limits_{(x,\,y)\in\mathbf{R}_2}\Phi(x,y)=\Phi(x^*,y^*)=0$, то $(x^*;\,y^*)$ — искомое решение системы (7.17), поскольку

$$\Phi(x^*,y^*)=0 \quad\Leftrightarrow\quad \begin{cases} f(x^*,y^*)=0, \\ g(x^*,y^*)=0. \end{cases}$$

Последовательность точек $(x_k;\,y_k)$ — приближений к точке $(x^*;\,y^*)$ минимума $\Phi(x,y)$ — обычно получают по рекуррентной формуле

$$\begin{pmatrix} x_{k+1} \\ y_{k+1}\end{pmatrix}=\begin{pmatrix} x_k \\ y_k\end{pmatrix}+\alpha_k\begin{pmatrix} p_k \\ q_k\end{pmatrix}, \quad k=0,1,2,\ldots, \qquad (7.55)$$

где $(p_k;\,q_k)^T$ — вектор, определяющий ***направление минимизации***, а α_k — скалярная величина, характеризующая величину шага минимизации (***шаговый множитель***). Учитывая геометрический смысл задачи минимизации функции двух переменных $\Phi(x,y)$ — «спуск на дно» поверхности $z=\Phi(x,y)$

*) Вообще говоря, не единственная.

(см. рис. 7.1), итерационный метод (7.55) можно назвать ***методом спуска***, если вектор $(p_k; q_k)^T$ при каждом k является ***направлением спуска*** (т.е. существует $\alpha > 0$ такое, что $\Phi(x_k + \alpha p_k, y_k + \alpha q_k) < \Phi(x_k, y_k)$) и если множитель α_k подбирается так, чтобы выполнялось ***условие релаксации*** $\Phi(x_{k+1}, y_{k+1}) < \Phi(x_k, y_k)$, означающее переход на каждой итерации в точку с меньшим значением минимизируемой функции.

Итак, при построении численного метода вида (7.55) минимизации функции $\Phi(x, y)$ следует ответить на два главных вопроса: как выбирать направление спуска $(p_k, q_k)^T$ и как регулировать длину шага в выбранном направлении с помощью скалярного параметра — шагового множителя α_k. Приведем наиболее простые соображения по этому поводу.

При выборе направления спуска естественным является выбор такого направления, в котором минимизируемая функция убывает наиболее быстро. Как известно из математического анализа функций нескольких переменных, направление наибольшего возрастания функции в данной точке показывает ее градиент в этой точке. Поэтому примем за направление спуска вектор

$$\begin{pmatrix} p_k \\ q_k \end{pmatrix} := -\operatorname{grad} \Phi(x_k, y_k) = -\begin{pmatrix} \Phi'_x(x_k, y_k) \\ \Phi'_y(x_k, y_k) \end{pmatrix}$$

— антиградиент функции $\Phi(x, y)$. Таким образом, из семейства методов (7.55) выделяем ***градиентный метод***

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} := \begin{pmatrix} x_k \\ y_k \end{pmatrix} - \alpha_k \begin{pmatrix} \Phi'_x(x_k, y_k) \\ \Phi'_y(x_k, y_k) \end{pmatrix}. \qquad (7.56)$$

Оптимальный шаг в направлении антиградиента — это такой шаг, при котором значение $\Phi(x_{k+1}, y_{k+1})$ — наименьшее среди всех других значений $\Phi(x, y)$ в этом фиксированном направлении, т.е. когда точка $(x_{k+1}; y_{k+1})$ является точкой условного минимума. Следовательно, можно рассчитывать на наиболее быструю сходимость метода (7.56), если полагать в нем

$$\alpha_k = \arg\min_{\alpha > 0} \Phi\big(x_k - \alpha \Phi'_x(x_k, y_k),\, y_k - \alpha \Phi'_y(x_k, y_k)\big). \qquad (7.57)$$

Такой выбор шагового множителя, называемый ***исчерпывающим спуском***, вместе с формулой (7.56) определяет ***метод наискорейшего спуска***.

Геометрическая интерпретация этого метода хорошо видна из рис. 7.1, 7.2. Характерны девяностоградусные изломы траектории наискорейшего спуска, что объясняется исчерпываемостью спуска и свойством градиента (а значит, и антиградиента) быть перпендикулярным к касательной к линии уровня в соответствующей точке *).

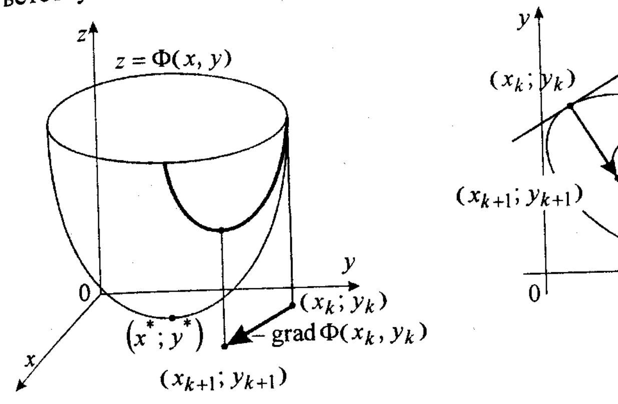


Рис. 7.1. Пространственная интерпретация метода наискорейшего спуска для функции (7.54)

Рис. 7.2. Траектория наискорейшего спуска для функции (7.54)

Наиболее типичной является ситуация, когда найти точно (аналитическими методами) оптимальное значение α_k не удается. Следовательно, приходится делать ставку на применение каких-либо численных методов одномерной минимизации и находить α_k в (7.57) лишь приближенно.

Несмотря на то, что задача нахождения минимума функции одной переменной $\varphi_k(\alpha) = \Phi\big(x_k - \alpha \Phi'_x(x_k, y_k), y_k - \alpha \Phi'_y(x_k, y_k)\big)$ намного проще, чем решаемая задача, применение тех или иных численных методов нахождения значений $\alpha_k = \arg\min \varphi_k(\alpha)$ с той или иной точностью требует вычисления нескольких значе-

---

*) Представив образно процесс движения текущей точки по траектории наискорейшего спуска на участке от положения $(x_k; y_k)$ до $(x_{k+1}; y_{k+1})$, можно отметить, что спуск характеризуется пересечением линий уровня $\Phi(x, y) = c$ все с меньшими значениями c до тех пор, пока не произойдет касание некоторой линии уровня (сплошь заполняющих $D(\Phi)$); дальнейшее движение в этом направлении приведет к пересечению линий уровня $\Phi(x, y) = c$ с увеличивающимися значениями параметра c.

ний минимизируемой функции. Так как это нужно делать на каждом итерационном шаге, то при большом числе шагов реализация метода наискорейшего спуска в чистом виде является достаточно высокозатратной. Существуют эффективные схемы приближенного вычисления квазиоптимальных α_k, в которых учитывается специфика минимизируемых функций (типа сумм квадратов функций) [61].

Зачастую успешной является такая стратегия градиентного метода, при которой шаговый множитель α_k в (7.56) берется либо сразу достаточно малым постоянным, либо предусматривается его уменьшение, например, делением пополам для удовлетворения условию релаксации на очередном шаге. Хотя каждый отдельный шаг градиентного метода при этом, вообще говоря, далек от оптимального, такой процесс по количеству вычислений функции может оказаться более эффективным, чем метод наискорейшего спуска.

Главное достоинство градиентных методов решения нелинейных систем — глобальная сходимость. Нетрудно доказать, что процесс градиентного спуска приведет к какой-либо точке минимума функции из любой начальной точки. При определенных условиях найденная точка минимума будет искомым решением исходной нелинейной системы.

Главный недостаток — медленная сходимость. Доказано, что сходимость таких методов — лишь линейная, причем, если для многих методов, таких как метод Ньютона, характерно ускорение сходимости при приближении к решению, то здесь имеет место скорее обратное. Поэтому есть резон в построении гибридных алгоритмов, которые начинали бы поиск искомой точки — решения данной нелинейной системы — глобально сходящимся градиентным методом, а затем производили уточнение каким-то быстросходящимся методом, например, методом Ньютона (разумеется, если данные функции обладают нужными свойствами).

Разработан ряд методов решения экстремальных задач, которые соединяют в себе низкую требовательность к выбору начальной точки и высокую скорость сходимости. К таким методам, называемым ***квазиньютоновскими***, можно отнести, например, ***метод переменной метрики (Дэвидона-Флетчера-Пауэлла)***, ***симметричный*** и ***положительно определенный методы секущих*** (на основе формулы пересчета Бройдена, см. § 7.4), а также уже упоминавшийся ранее применительно к СЛАУ (см. гл. 3) ***метод сопряженных градиентов***.

При наличии негладких функций в решаемой задаче следует отказаться от использования производных или их аппроксимаций и прибегнуть к так называемым ***методам прямого поиска*** (***циклического покоординатного спуска***, ***Хука и Дживса***, ***Розенброка*** и т.п.). Описание упомянутых и многих других методов такого типа можно найти в учебной и в специальной литературе, посвященной решению экстремальных задач (см., например, [32, 49, 52, 68, 108]).

Замечание 7.4. Для разных семейств численных методов минимизации могут быть рекомендованы свои критерии останова итерационного процесса. Например, учитывая, что в точке минимума дифференцируемой функции должно выполняться необходимое условие экстремума, на конец счета градиентным методом можно выходить, когда станет достаточно малой норма градиента. Если же принять во внимание, что минимизация применяется к решению нелинейной системы, то целесообразно отслеживать близость к нулю значений минимизируемой неотрицательной функции $\Phi(x, y)$, т.е. судить о точности получаемого приближения по квадрату его евклидовой нормы невязки.

Замечание 7.5. Как отмечалось в начале этого параграфа, ограничение размерности решаемой системы здесь делалось сугубо из иллюстративных соображений. Ничто не помешает развить рассматриваемый подход на случай n-мерной системы (7.1), сводя ее решение к экстремальной задаче

$$\Phi(\mathbf{x}) := \sum_{i=1}^{n} f_i^2(\mathbf{x}) \to \min .$$

7.7. ЧИСЛЕННЫЙ ПРИМЕР

Типичное поведение рассмотренных методов решения систем нелинейных уравнений отражает следующий численный пример.

Пусть ищется решение системы

$$\begin{cases} 20\ln(x-y) - x - y - 6 = 0, \\ 20\sin(0.7x - 0.7y) + 7x + 7y = 0 \end{cases} \quad (7.58)$$

в окрестности точки $x_0 = 0$, $y_0 = -1$.

Результаты применения к этой системе разных описанных выше итерационных процессов, начинающихся с данной точки $(x_0; y_0)$ и заканчивающихся, как только выполнится неравенство

$$\max\{|x_k - x_{k-1}|, |y_k - y_{k-1}|\} < \varepsilon, \quad (7.59)$$

где k —номер итерации, представлены в табл. 7.1. В ней приведены: приближенные решения, полученные на k-й итерации с

помощью девяти перечисленных там методов, значения k, при которых сработал критерий (7.59) с $\varepsilon = 0.000001$, а также векторы невязок, характеризующие некоторую меру близости указанного приближения к точному решению системы (7.58). При этом всюду шаг аппроксимации производных (начальный шаг в методах секущих и Брауна) брался равным ε в каждой компоненте.

Таблица 7.1

Результаты решения системы (7.58) *разными методами*

№ п/п	Метод	Приближенное решение $(x_k; y_k)^T$	Число итераций k	Невязка $(f(x_k, y_k); g(x_k, y_k))^T$
1	Ньютона	–0.46584782 –1.67846886	4	0.000000000001 –0.000000000002
2	Разностный Ньютона	–0.46584782 –1.67846886	4	–0.000000000497 –0.000000000261
3	Простейший секущих	–0.46584782 –1.67846886	5	0.000000000001 –0.000000000002
4	Ньютона с аппроксимацией обратных матриц	–0.46584782 –1.67846886	4	0.000000000098 –0.000000000048
5	Брауна (с аппроксимацией производных)*)	–0.46584782 –1.67846886	5	–0.000000000000 –0.000000037367
6	Полюсный Ньютона**)	–0.46584782 –1.67846886	4	–0.000000000000 –0.000000000000
7	Секущих Бройдена	–0.46584781 –1.67846886	5	0.000000100053 0.000000040792
8	Модифицированный (упрощенный) Ньютона	–0.46584784 –1.67846880	8	–0.000001418985 –0.000000589060
9	Наискорейшего спуска	–0.46574667 –1.67846711	36	–0.000012735642 0.000014735135

Число итераций и, соответственно, число нулей перед первыми значащими цифрами в невязках, подкрепляет сделанные выше заявления о быстрой сходимости методов 1–7 и справедливость высказанных в замечании 7.2 соображений о надежности и даже некоторой грубости простого критерия останова (7.16) (для

*) Аппроксимация производных на k-й итерации осуществлялась с шагом $h_k = \min\{|p_k|, |q_k|\}$.

**) С постоянной матрицей $\mathbf{C} := \begin{pmatrix} 1 & 2 \\ 2 & 0 \end{pmatrix}$ и переменным вектором $\mathbf{d} := F(\mathbf{x}^{(k)})$, т.е. с полюсами $P_1^{(k)} = (1; 2; f(x_k, y_k))$, $P_2^{(k)} = (2; 0; g(x_k, y_k))$.

системы (7.58) реализованного в виде (7.59)). Сравнивая шестые знаки после запятой у приближенных решений, полученных методом наискорейшего спуска и, например, методом Ньютона, видим, что для медленно сходящихся методов примененный критерий уже не является столь надежным (из процесса итерирования в градиентных методах, как уже упоминалось в замечании 7.4, обычно выходят по своим критериям, например, по малости нормы градиента).

7.8. СХОДИМОСТЬ МЕТОДА НЬЮТОНА И НЕКОТОРЫХ ЕГО МОДИФИКАЦИЙ

Многие утверждения о сходимости метода Ньютона восходят к известным результатам Л.В. Канторовича*), перенесшего метод Ньютона на нелинейные операторные уравнения в банаховых пространствах, в связи с чем и метод в таком общем случае часто называют ***методом Ньютона–Канторовича***. В основе этих результатов лежит принцип мажорирования операторного уравнения некоторым скалярным уравнением, через корни которого оценивается абсолютная погрешность приближений, получаемых в итерационном процессе (см. [80, 129] и др.). Таким путем получают условия квадратичной сходимости основного (7.6) и линейной сходимости модифицированного (упрощенного) (7.8) методов Ньютона. Непосредственное применение подобной методики к анализу сходимости методов более общего вида, в частности, ААМН (7.11), затруднительно. Поэтому изберем другой путь исследования сходимости итерационных последовательностей. Вообще говоря, здесь будет эксплуатироваться та же идея скалярного мажорирования, только в несколько ином виде.

Прежде всего, попытаемся выяснить, какие условия нужно наложить на скалярную последовательность (p_k), мажорирующую последовательности норм поправок $\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}$ и невязок $F(\mathbf{x}^{(k)})$, чтобы последовательность n-мерных векторов $\mathbf{x}^{(k)}$ сходилась к нулю вектор-функции $F(\mathbf{x})$ с заданным порядком μ (≥ 1) скорости сходимости независимо от способа получения элементов последовательности $(\mathbf{x}^{(k)})$.

*) Канторович Леонид Витальевич (1912–1986) — российский академик, приобретший мировую известность своими основополагающими работами в области линейного программирования и применения функционального анализа в вычислительной математике.

Теорема 7.3. **I.** *Пусть непрерывная векторная функция* $F : M \subseteq \mathbf{R}_n \to \mathbf{R}_n$ *) *и последовательность векторов* $\mathbf{x}^{(k)} \in M$ *таковы, что при всех* $k \in \mathbf{N}_0$ *выполняются условия:*

$$1)\ \left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \le \lambda p_k; \quad 2)\ \left\|F\left(\mathbf{x}^{(k)}\right)\right\| \le p_k,$$

где числа p_k *определяются рекуррентным равенством*

$$p_{k+1} = G_0 p_k^{\mu}, \quad k = 0, 1, 2, \ldots, \tag{7.60}$$

а $\lambda > 0$, $G_0 > 0$, $p_0 > 0$ *и* $\mu > 1$ *— некоторые числовые параметры.*

II. *Тогда, если* $\nu := G_0 p_0^{\mu-1} < 1$ *и замкнутый шар* $S\left(\mathbf{x}^{(0)},\ r := \lambda\, p_0 \sum_{i=0}^{\infty} \nu^{\frac{\mu^i - 1}{\mu - 1}}\right)$ *содержится в* M, *то все члены последовательности* $\left(\mathbf{x}^{(k)}\right)$ *принадлежат* S, *последовательность* $\left(\mathbf{x}^{(k)}\right)$ *имеет предел* $\mathbf{x}^* \in S$ *такой, что* $F\left(\mathbf{x}^*\right) = \mathbf{0}$; *при этом быстрота сходимости* $\left(\mathbf{x}^{(k)}\right)$ *к* $\mathbf{x}^*$ *характеризуется неравенством* $(\forall\, k \in \mathbf{N})$

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \frac{\lambda p_0}{1 - \nu^{\mu^k}} \nu^{\frac{\mu^k - 1}{\mu - 1}}. \tag{7.61}$$

Доказательство. Пользуясь равенством (7.60), выразим элементы последовательности (p_i) через ее начальный член p_0 и определенную в теореме величину ν:

$$p_i = G_0 p_{i-1}^{\mu} = G_0 \left(G_0 p_{i-2}^{\mu}\right)^{\mu} = G_0^{1+\mu} p_{i-2}^{\mu^2} = G_0^{1+\mu+\mu^2} \cdot p_{i-3}^{\mu^3} = \ldots =$$
$$= G_0^{1+\mu+\mu^2+\ldots+\mu^{i-1}} p_0^{\mu^i} = G_0^{\frac{\mu^i-1}{\mu-1}} \cdot p_0^{\mu^i - 1} \cdot p_0^{1-\mu^i} \cdot p_0^{\mu^i} = p_0 \cdot \nu^{\frac{\mu^i - 1}{\mu - 1}}.$$

Следовательно, условие 1) доказываемой теоремы можно пере-

*) Без всяких прочих изменений можно заменить здесь $\mathbf{R}_n \to \mathbf{R}_n$ на более общий случай $\mathbf{R}_n \to \mathbf{R}_m$.

писать в виде

$$\left\|\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\right\| \le \lambda p_0 \cdot \nu^{\frac{\mu^i - 1}{\mu - 1}}. \tag{7.62}$$

Посредством (7.62) теперь устанавливаем, что

$$\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(0)}\right\| \le \sum_{i=0}^{k} \left\|\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\right\| \le \sum_{i=0}^{k} \lambda p_0 \cdot \nu^{\frac{\mu^i - 1}{\mu - 1}} \le r,$$

т.е. все члены заданной последовательности $\left(\mathbf{x}^{(k)}\right)$ принадлежат $S \subseteq M$. Покажем, что она удовлетворяет критерию Коши. С помощью того же неравенства (7.62) имеем:

$$\left\|\mathbf{x}^{(k+m)} - \mathbf{x}^{(k)}\right\| \le \sum_{i=k}^{k+m-1} \left\|\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\right\| \le \lambda p_0 \sum_{i=k}^{k+m-1} \nu^{\frac{\mu^i - 1}{\mu - 1}} =$$
$$= \lambda p_0 \cdot \nu^{\frac{\mu^k - 1}{\mu - 1}} \left(1 + \nu^{\mu^k} + \nu^{\mu^k + \mu^{k+1}} + \ldots + \nu^{\mu^k + \mu^{k+1} + \ldots + \mu^{k+m-1}}\right) <$$
$$< \lambda p_0 \cdot \nu^{\frac{\mu^k - 1}{\mu - 1}} \left(1 + \nu^{\mu^k} + \nu^{2\mu^k} + \ldots + \nu^{(m-1)\mu^k}\right) = \lambda p_0 \cdot \nu^{\frac{\mu^k - 1}{\mu - 1}} \cdot \frac{1 - \nu^{m\mu^k}}{1 - \nu^{\mu^k}}.$$

Полученное неравенство, рассматриваемое при фиксированном $m \in \mathbf{N}$ и $k \to \infty$, говорит о фундаментальности $\left(\mathbf{x}^{(k)}\right)$ и существовании предельного вектора $\mathbf{x}^*$ в шаре S (в силу предполагаемой замкнутости S). С другой стороны, если в нем зафиксировать k и перейти к пределу при $m \to \infty$, сразу получается утверждаемая оценка (7.61). Подстановка выражения $p_k = p_0 \cdot \nu^{\frac{\mu^k - 1}{\mu - 1}}$ в условие 2) показывает, что $\left\|F\left(\mathbf{x}^{(k+1)}\right)\right\| \to 0$ при $k \to \infty$ (т.е. при $\mathbf{x}^{(k)} \to \mathbf{x}^*$), а это, в силу предполагаемой непрерывности $F(\mathbf{x})$, означает, что $\mathbf{x}^* := \lim_{k \to \infty} \mathbf{x}^{(k)}$ есть решение уравнения $F(\mathbf{x}) = \mathbf{0}$. Теорема доказана.

Замечание 7.6. Нетрудно убедиться, что теорема 7.3 (а также следующая теорема 7.4) справедлива и при $\mu = 1$. При этом

$$\nu := G_0, \quad r := \frac{\lambda p_0}{1 - \nu}, \quad \left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \frac{\lambda p_0}{1 - \nu^k} \cdot \nu^k$$

(т.е. сходимость $\left(\mathbf{x}^{(k)}\right)$ к $\mathbf{x}^*$ в шаре $S\left(\mathbf{x}^{(0)}, r\right)$ в этом случае — линейная).

Изменим требования к последовательности $\left(\mathbf{x}^{(k)}\right)$ и к вектор-функции $F(\mathbf{x})$, фигурирующие в части **I** теоремы 7.3, так, чтобы осталась неизменной ее констатирующая часть **II**.

Теорема 7.4. *Пусть существуют такие последовательности положительных чисел* H_k *и* G_k, *удовлетворяющих условиям* $H_k \le H_0$, $G_k \le G_0$, *и число* $\mu > 1$, *что в предположении, что* $\mathbf{x}^{(k)} \in M$, *при всех* $k \in N_0$ *выполняются неравенства*:

1) $\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \le H_k \cdot \left\|F\left(\mathbf{x}^{(k)}\right)\right\|$;

2) $\left\|F\left(\mathbf{x}^{(k+1)}\right)\right\| \le G_k \cdot \left\|F\left(\mathbf{x}^{(k)}\right)\right\|^{\mu}$.

Тогда справедлива часть ***II*** *теоремы* 7.3 *с* $p_0 \ge \left\|F\left(\mathbf{x}^{(0)}\right)\right\|$, $\lambda := H_0$.

Доказательство. Определим последовательность (p_k) равенством $p_{k+1} = G_0 p_k^{\mu}$ и по индукции покажем, что эта последовательность мажорирует $\left(\left\|F\left(\mathbf{x}^{(k)}\right)\right\|\right)$ одновременно с доказательством принадлежности векторов $\mathbf{x}^{(k)}$ шару $S\left(\mathbf{x}^{(0)}, r\right)$.

По условию $\mathbf{x}^{(0)} \in S$ и $\left\|F\left(\mathbf{x}^{(0)}\right)\right\| \le p_0$. Сделаем индукционное предположение, что

$$\mathbf{x}^{(i)} \in S \quad \text{и} \quad \left\|F\left(\mathbf{x}^{(i)}\right)\right\| \le p_i \quad \forall i \in \{0, 1, \dots, k\}.$$

Тогда

$$\left\|\mathbf{x}^{(i+1)} - \mathbf{x}^{(i)}\right\| \le H_i \cdot \left\|F\left(\mathbf{x}^{(i)}\right)\right\| \le H_0 p_i = H_0 p_0 \cdot \nu^{\frac{\mu^i - 1}{\mu - 1}}.$$

Из этого следует (см. доказательство теоремы 7.3) неравенство $\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(0)}\right\| \le r$, означающее, что $\mathbf{x}^{(k+1)} \in S \subseteq M$. Таким образом, значение $F\left(\mathbf{x}^{(k+1)}\right)$ существует и

$$\left\|F\left(\mathbf{x}^{(k+1)}\right)\right\| \le G_k \left\|F\left(\mathbf{x}^{(k)}\right)\right\|^{\mu} \le G_0 p_k^{\mu} = p_{k+1}.$$

Теперь можно сказать, что условия теоремы 7.3 полностью выполнены, значит справедливо и ее заключение.

Прежде чем применить доказанные выше теоремы к конкретным методам типа (7.5), выведем из формулы Тейлора*)

$$F(\mathbf{x}) \equiv F\left(\mathbf{x}^{(0)} + \mathbf{h}\right) = F\left(\mathbf{x}^{(0)}\right) + F'\left(\mathbf{x}^{(0)}\right)\mathbf{h} + \frac{1}{2!}F''\left(\mathbf{x}^{(0)}\right)\mathbf{h}^2 + $$

$$+ \ldots + \frac{1}{(l-1)!}F^{(l-1)}\left(\mathbf{x}^{(0)}\right)\mathbf{h}^{l-1} + \omega\left(\mathbf{x}^{(0)}, \mathbf{h}\right) \quad (7.63)$$

с остаточным членом

$$\omega\left(\mathbf{x}^{(0)}, \mathbf{h}\right) = \frac{1}{(l-1)!}\int_{\mathbf{x}^{(0)}}^{\mathbf{x}} F^{(l)}(\mathbf{z})(\mathbf{x} - \mathbf{z})^{l-1} d\mathbf{z} \quad (7.64)$$

простое неравенство для оценивания $\|F(\mathbf{x})\|$ в произвольной точке $\mathbf{x} \in M$ через значения F и F' в близкой к $\mathbf{x}$ точке $\mathbf{x}^{(0)} \in M$.

Лемма 7.2. *Пусть векторная функция* $F(\mathbf{x})$ *в области* $M \subseteq \mathbf{R}_n$ *дифференцируема по Фреше и ее производная удовлетворяет* ***условию Липшица***:

$$\left\|F'(\mathbf{x}) - F'\left(\mathbf{x}^{(0)}\right)\right\| \le L \cdot \left\|\mathbf{x} - \mathbf{x}^{(0)}\right\| \quad \forall \mathbf{x}, \mathbf{x}^{(0)} \in M.$$

Тогда при любых $\mathbf{x}$, $\mathbf{x}^{(0)}$ *из* M

$$\|F(\mathbf{x})\| \le \left\|F\left(\mathbf{x}^{(0)}\right) + F'\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x} - \mathbf{x}^{(0)}\right)\right\| + \frac{L}{2}\left\|\mathbf{x} - \mathbf{x}^{(0)}\right\|^2. \quad (7.65)$$

Доказательство. Запишем формулу Тейлора (7.63) с остаточным членом (7.64) для случая $l = 1$:

$$F(\mathbf{x}) = F\left(\mathbf{x}^{(0)}\right) + \int_{\mathbf{x}^{(0)}}^{\mathbf{x}} F'(\mathbf{z}) d\mathbf{z}.$$

Учитывая, что $\int_{\mathbf{x}^{(0)}}^{\mathbf{x}} F'\left(\mathbf{x}^{(0)}\right) d\mathbf{z} \equiv F'\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x} - \mathbf{x}^{(0)}\right)$, ее можно преобразовать к виду

$$F(\mathbf{x}) = F\left(\mathbf{x}^{(0)}\right) + F'\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x} - \mathbf{x}^{(0)}\right) + \int_{\mathbf{x}^{(0)}}^{\mathbf{x}}\left[F'(\mathbf{z}) - F'\left(\mathbf{x}^{(0)}\right)\right] d\mathbf{z}.$$

---

*) Выражения типа $F^{(i)}\mathbf{h}^i$ следует понимать как результат применения i-линейного оператора i-кратного дифференцирования $F^{(i)}$ к вектору $\mathbf{h}$.

В последнем представлении $F(\mathbf{x})$ интеграл по отрезку $\left[\mathbf{x}^{(0)}, \mathbf{x}\right]$ заменой $\mathbf{z} = \mathbf{x}^{(0)} + \tau\left(\mathbf{x} - \mathbf{x}^{(0)}\right)$ сведем к интегралу по абстрактной переменной $\tau \in [0, 1]$. Имеем:

$$F(\mathbf{x}) = F\left(\mathbf{x}^{(0)}\right) + F'\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x} - \mathbf{x}^{(0)}\right) +$$

$$+ \int_0^1 \left[F'\left(\mathbf{x}^{(0)} + \tau\left(\mathbf{x} - \mathbf{x}^{(0)}\right)\right) - F'\left(\mathbf{x}^{(0)}\right)\right]\left(\mathbf{x} - \mathbf{x}^{(0)}\right) d\tau.$$

Отсюда, переходя к нормам, получаем доказываемое неравенство (7.65), предварительно оценив норму интеграла следующим образом:

$$\left\| \int_0^1 \left[F'\left(\mathbf{x}^{(0)} + \tau\left(\mathbf{x} - \mathbf{x}^{(0)}\right)\right) - F'\left(\mathbf{x}^{(0)}\right)\right]\left(\mathbf{x} - \mathbf{x}^{(0)}\right) d\tau \right\| \le$$

$$\le \int_0^1 \left\| F'\left(\mathbf{x}^{(0)} + \tau\left(\mathbf{x} - \mathbf{x}^{(0)}\right)\right) - F'\left(\mathbf{x}^{(0)}\right) \right\| \cdot \left\| \mathbf{x} - \mathbf{x}^{(0)} \right\| d\tau \le$$

$$\le \int_0^1 L \left\| \mathbf{x} - \mathbf{x}^{(0)} \right\|^2 \tau d\tau = \frac{L}{2} \left\| \mathbf{x} - \mathbf{x}^{(0)} \right\|^2.$$

Теорема 7.5. *Пусть функция $F(\mathbf{x})$ определена и дифференцируема по Фреше в некоторой открытой области $M \subseteq \mathbf{R}_n$, причем:*

1) $\exists L > 0: \ \left\|F'(\mathbf{x}) - F'(\widetilde{\mathbf{x}})\right\| \le L\left\|\mathbf{x} - \widetilde{\mathbf{x}}\right\| \quad \forall \mathbf{x}, \widetilde{\mathbf{x}} \in M$; *)

2) $\exists [F'(\mathbf{x})]^{-1}$ *и* $\exists C > 0: \left\|[F'(\mathbf{x})]^{-1}\right\| \le C \quad \forall \mathbf{x} \in M$.

Тогда, если

$$\nu := 0.5 L C^2 p_0 < 1, \quad \text{где} \quad p_0 \ge \left\| F\left(\mathbf{x}^{(0)}\right) \right\|,$$

и замкнутый шар $S\left(\mathbf{x}^{(0)}, \ r := C p_0 \sum_{i=0}^{\infty} \nu^{2^i - 1}\right)$ *целиком содержится в M, то все члены последовательности $\left(\mathbf{x}^{(k)}\right)$, определяемые методом Ньютона (7.6), начинающимся с заданного $\mathbf{x}^{(0)}$, лежат в $S \subseteq M$; последовательность $\left(\mathbf{x}^{(k)}\right)$ имеет предел $\mathbf{x}^* \in S$, служащий решением уравнения $F(\mathbf{x}) = \mathbf{0}$; справедлива оценка погрешности*

$$\left\| \mathbf{x}^* - \mathbf{x}^{(k)} \right\| \le \frac{C p_0}{1 - \nu^{2^k}} \cdot \nu^{2^k - 1}.$$

Доказательство. Непосредственно из равенства (7.6) получаем

$$\left\| \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \right\| \le \left\| \left[F'\left(\mathbf{x}^{(k)}\right)\right]^{-1} \right\| \cdot \left\| F\left(\mathbf{x}^{(k)}\right) \right\| \le C \cdot \left\| F\left(\mathbf{x}^{(k)}\right) \right\|,$$

т.е. требование 1) теоремы 7.4 с $H_k \equiv C$.

Далее обратимся к неравенству (7.65), установленному леммой 7.2. Положив в нем $\mathbf{x} := \mathbf{x}^{(k+1)}$, $\mathbf{x}^{(0)} := \mathbf{x}^{(k)}$, приведем (7.65) к виду

$$\left\| F\left(\mathbf{x}^{(k+1)}\right) \right\| \le \left\| F\left(\mathbf{x}^{(k)}\right) + F'\left(\mathbf{x}^{(k)}\right)\left(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right) \right\| + \frac{L}{2} \left\| \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \right\|^2. \quad (7.66)$$

Но в данном случае, т.е. для метода Ньютона,

$$F\left(\mathbf{x}^{(k)}\right) + F'\left(\mathbf{x}^{(k)}\right)\left(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right) = \mathbf{0},$$

поэтому

$$\left\| F\left(\mathbf{x}^{(k+1)}\right) \right\| \le \frac{L}{2} \left\| \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} \right\|^2 \le \frac{1}{2} L C^2 \left\| F\left(\mathbf{x}^{(k)}\right) \right\|^2.$$

Таким образом, выполнено и требование 2) теоремы 7.4 с $G_k = \frac{1}{2} L C^2$, $\mu = 2$.

Завершает доказательство подстановка постоянных $\lambda = H_0 = C$, $G_0 = \frac{1}{2} L C^2$, $\mu = 2$ в часть **II** теоремы 7.3.

Для модифицированного метода Ньютона требование обратимости матрицы Якоби в любой точке M заменим менее ограничительным требованием ее обратимости лишь в начальной точке $\mathbf{x}^{(0)}$.

Теорема 7.6. *Пусть для* $F: \left(M \subseteq \mathbf{R}_n\right) \to \mathbf{R}_n$:

$\exists F'(\mathbf{x}): (\exists L > 0: \left\|F'(\mathbf{x}) - F'(\widetilde{\mathbf{x}})\right\| \le L\left\|\mathbf{x} - \widetilde{\mathbf{x}}\right\| \ \forall \widetilde{\mathbf{x}} \in M) \ \forall \mathbf{x} \in M$;

$\exists \left[F'\left(\mathbf{x}^{(0)}\right)\right]^{-1}, \quad \exists C_0 > 0: \left\| \left[F'\left(\mathbf{x}^{(0)}\right)\right]^{-1} \right\| \le C_0.$

*) Согласно лемме 7.1, для дважды непрерывно дифференцируемой функции $F(\mathbf{x})$ в качестве L можно брать оценку сверху величины $\left\|F''(\mathbf{x})\right\|$.

Тогда, если при $p_0 \ge \left\| F\left(\mathbf{x}^{(0)}\right)\right\|$ *величина*

$$t := LC_0^2 p_0 \le 0.125 \tag{7.67}$$

и замкнутый шар $S\left(\mathbf{x}^{(0)},\ r := \dfrac{2C_0 p_0}{1 \pm \sqrt{1-8t}}\right)$ *содержится в* M*, то начатый с* $\mathbf{x}^{(0)}$ *модифицированный метод Ньютона* (7.8) *сходится в* S *к решению* $\mathbf{x}^*$ *уравнения* (7.1а) *с оценкой погрешности*

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \frac{C_0 p_0}{1-\nu^k}\cdot \nu^k,$$

*где**) $\nu := \dfrac{1}{2} \mp \sqrt{\dfrac{1}{4} - 2t}$.

Доказательство. Как и в предыдущей теореме, требование 1) теоремы 7.4 получается сразу же из формулы, определяющей метод:

$$\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \le \left\| \left[F'\left(\mathbf{x}^{(0)}\right)\right]^{-1} F\left(\mathbf{x}^{(k)}\right)\right\| \le C_0 \left\|F\left(\mathbf{x}^{(k)}\right)\right\|.$$

Для оценки $\left\|F\left(\mathbf{x}^{(k+1)}\right)\right\|$ преобразуем неравенство (7.66) так, чтобы воспользоваться равенством нулю выражения $F\left(\mathbf{x}^{(k)}\right) + F'\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right)$ в соответствии с (7.8). Имеем:

$$\left\|F\left(\mathbf{x}^{(k+1)}\right)\right\| \le \Big\|F\left(\mathbf{x}^{(k)}\right) + F'\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right) +$$

$$+ \left[F'\left(\mathbf{x}^{(k)}\right) - F'\left(\mathbf{x}^{(0)}\right)\right]\left(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right)\Big\| + \frac{L}{2}\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|^2 \le$$

$$\le \left(\left\|F'\left(\mathbf{x}^{(k)}\right) - F'\left(\mathbf{x}^{(0)}\right)\right\| + \frac{L}{2}\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|\right)\cdot\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \le$$

$$\le L\left(\left\|\mathbf{x}^{(k)} - \mathbf{x}^{(0)}\right\| + \frac{1}{2}\left(\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(0)}\right\| + \left\|\mathbf{x}^{(0)} - \mathbf{x}^{(k)}\right\|\right)\right)\cdot C_0 \cdot \left\|F\left(\mathbf{x}^{(k)}\right)\right\| \le$$

$$\le LC_0\left(r + \frac{1}{2}(r+r)\right)\cdot\left\|F\left(\mathbf{x}^{(k)}\right)\right\| = 2LC_0 r\cdot\left\|F\left(\mathbf{x}^{(k)}\right)\right\|.$$

---

*) В выражениях r и ν одновременно берутся либо только верхние, либо только нижние знаки.

Отсюда видно, что можно считать выполненным требование 2) теоремы 7.4 с $G_k = G_0 = 2LC_0 r$ и $\mu = 1$.

Подстановка постоянных $\mu = 1$, $\lambda = C_0$ в заключительную часть **II** теоремы 7.3 показывает (с учетом замечания 7.5), что в данном случае должно быть $r = \dfrac{C_0 p_0}{1-\nu}$, а $\nu = 2LC_0 r < 1$. Исключая из последних двух равенств параметр r, получаем квадратное относительно ν уравнение

$$\nu^2 - \nu + 2t = 0,$$

оба корня которого $\nu_{1,2} = 0.5 \mp \sqrt{0.25 - 2t}$ положительны и меньше 1, если только неотрицателен дискриминант $0.25 - 2t$, что обеспечивается условием (7.67). Подставляя эти значения ν, находим связанные с ним значения радиуса r шара S. Теперь справедливость заключения данной теоремы очевидна.

Обратимся, наконец, к обоснованию квадратичной сходимости метода Ньютона с последовательной аппроксимацией обратных матриц, т.е. ААМН (7.11).

Теорема 7.7. *Пусть функция* $F(\mathbf{x})$ *определена и дифференцируема в* $M \subseteq \mathbf{R}_n$*, причем*

$$\exists L > 0 : \left\|F'(\mathbf{x}) - F'(\tilde{\mathbf{x}})\right\| \le L\left\|\mathbf{x} - \tilde{\mathbf{x}}\right\| \quad \forall \mathbf{x}, \tilde{\mathbf{x}} \in M.$$

Тогда, если вектор $\mathbf{x}^{(0)}$ *и матрица* $\mathbf{A}_0$ *таковы, что при некотором* $\lambda > 0$ *выполняются неравенства*

$$\left\|\mathbf{E} - F'\left(\mathbf{x}^{(0)}\right)\mathbf{A}_0\right\| \le L\lambda^2\left\|F\left(\mathbf{x}^{(0)}\right)\right\|, \tag{7.68}$$

$$\nu := 4L\lambda^2\left\|F\left(\mathbf{x}^{(0)}\right)\right\| \le 1 - \frac{\left\|\mathbf{A}_0\right\|}{2\lambda - \left\|\mathbf{A}_0\right\|} \tag{7.69}$$

и

$$S := \left\{\mathbf{x} \in R_n \,\middle|\, \left\|\mathbf{x} - \mathbf{x}^{(0)}\right\| \le \lambda\left\|F\left(\mathbf{x}^{(0)}\right)\right\|\cdot\sum_{i=0}^{\infty}\nu^i\right\} \subseteq M,$$

то начатый с данных $\mathbf{x}^{(0)}$, $\mathbf{A}_0$ *ААМН* (7.11) *сходится в* S *к решению* $\mathbf{x}^*$ *уравнения* $F(\mathbf{x}) = \mathbf{0}$ *и имеет место оценка погрешности*

$$\left\|\mathbf{x}^* - \mathbf{x}^{(k)}\right\| \le \frac{\lambda\left\|F\left(\mathbf{x}^{(0)}\right)\right\|}{1-\nu^{2^k}}\cdot\nu^{2^k}. \tag{7.70}$$

Доказательство. Введем в рассмотрение последовательности положительных величин p_k, β_k, b_k, определяемых при $k = 0, 1, 2, \ldots$ равенствами

$$p_{k+1} = 4L\lambda^2 p_k, \quad p_0 := \left\|F\left(\mathbf{x}^{(0)}\right)\right\|; \tag{7.71}$$

$$\beta_k = 2L\lambda^2 p_k; \tag{7.72}$$

$$b_{k+1} = \beta_k^2, \quad b_0 := L\lambda^2 p_0. \tag{7.73}$$

Очевидно невозрастание этих последовательностей. Легко также видеть, что

$$b_k = L\lambda^2 p_k \quad \forall\, k \in \mathbf{N}_0. \tag{7.74}$$

Действительно, предположив равенство (7.74) верным при некотором k, имеем

$$b_{k+1} = \left(2L\lambda^2 p_k\right)^2 = L\lambda^2 \cdot 4L\lambda^2 p_k^2 = L\lambda^2 p_{k+1},$$

т.е. то, что получили бы формальной заменой в (7.74) k на $k+1$.

Обозначим $\mathbf{B}_k := \mathbf{E} - F'\left(\mathbf{x}^{(k)}\right)\mathbf{A}_k$ (— невязка для $\mathbf{A}_k$ относительно $\left[F'\left(\mathbf{x}^{(k)}\right)\right]^{-1}$).

Докажем, что при любом $k \in \mathbf{N}_0$ скалярные последовательности (p_k), (β_k), (b_k) мажорируют последовательности норм векторов $F\left(\mathbf{x}^{(k)}\right)$ и матриц $\mathbf{\Psi}_k$, $\mathbf{B}_k$ соответственно и, вместе с тем, λ ограничивает сверху $\|\mathbf{A}_k\|$. С этой целью сделаем индукционное предположение, что одновременно выполняются неравенства:

$$\left\|F\left(\mathbf{x}^{(k)}\right)\right\| \le p_k, \quad \|\mathbf{\Psi}_{k-1}\| \le \beta_{k-1}, \quad \|\mathbf{B}_k\| \le b_k \quad \text{и} \quad \|\mathbf{A}_k\| \le \lambda. \tag{7.75}$$

Тогда можно проделать следующие выкладки:

$$\|\mathbf{\Psi}_k\| = \left\|\mathbf{E} - F'\left(\mathbf{x}^{(k+1)}\right)\mathbf{A}_k\right\| = \left\|\mathbf{B}_k + F'\left(\mathbf{x}^{(k)}\right)\mathbf{A}_k - F'\left(\mathbf{x}^{(k+1)}\right)\mathbf{A}_k\right\| \le$$
$$\le \|\mathbf{B}_k\| + L\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\| \cdot \|\mathbf{A}_k\| \le b_k + L\lambda\left\|\mathbf{A}_k F\left(\mathbf{x}^{(k)}\right)\right\| \le$$
$$\le L\lambda^2 p_k + L\lambda^2 p_k = \beta_k$$

(см. (7.11), (7.75), (7.74), (7.72));

$$\|\mathbf{B}_{k+1}\| = \left\|\mathbf{E} - F'\left(\mathbf{x}^{(k+1)}\right)\mathbf{A}_{k+1}\right\| = \left\|\mathbf{E} - F'\left(\mathbf{x}^{(k+1)}\right)\mathbf{A}_k - F'\left(\mathbf{x}^{(k+1)}\right)\mathbf{A}_k\mathbf{\Psi}_k\right\| =$$
$$= \left\|\mathbf{\Psi}_k^2\right\| \le \|\mathbf{\Psi}_k\|^2 \le \beta_k^2 = b_{k+1}$$

(см. (7.11), (7.73) и предыдущее неравенство);

$$\left\|F\left(\mathbf{x}^{(k+1)}\right)\right\| \le \left\|F\left(\mathbf{x}^{(k)}\right) - F'\left(\mathbf{x}^{(k)}\right)\mathbf{A}_k F\left(\mathbf{x}^{(k)}\right)\right\| + \frac{L}{2}\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|^2 \le$$
$$\le \|\mathbf{B}_k\| \cdot \left\|F\left(\mathbf{x}^{(k)}\right)\right\| + \frac{L}{2}\|\mathbf{A}_k\|^2 \cdot \left\|F\left(\mathbf{x}^{(k)}\right)\right\|^2 \le b_k p_k + \frac{L}{2}\lambda^2 p_k^2 =$$
$$= L\lambda^2 p_k^2 + \frac{L}{2}\lambda^2 p_k^2 < 4L\lambda^2 p_k^2 = p_{k+1}$$

(см. (7.66), (7.11), (7.75), (7.74), (7.71)); также из (7.11) следует, что

$$\|\mathbf{A}_{k+1}\| \le \|\mathbf{A}_k\|(1 + \|\mathbf{\Psi}_k\|) \le \|\mathbf{A}_{k-1}\|(1 + \|\mathbf{\Psi}_{k-1}\|)(1 + \|\mathbf{\Psi}_k\|) \le \ldots$$
$$\ldots \le \|\mathbf{A}_0\| \cdot \prod_{j=0}^{k}\left(1 + \left\|\mathbf{\Psi}_j\right\|\right) \le \|\mathbf{A}_0\| \cdot \prod_{j=0}^{k}\left(1 + \beta_j\right);$$

но $\beta_k = \frac{1}{2}\nu^{2^k}$ (действительно, из предположения, что $\beta_{k-1} = \frac{1}{2}\nu^{2^{k-1}}$, получаем

$$\beta_k = 2L\lambda^2 p_k = \frac{1}{2}\left(4L\lambda^2 p_{k-1}\right)^2 = \frac{1}{2}(2\beta_{k-1})^2 = \frac{1}{2}\left(\nu^{2^{k-1}}\right)^2 = \frac{1}{2}\nu^{2^k}),$$

поэтому далее можно продолжить оценивание так:

$$\|\mathbf{A}_{k+1}\| \le \|\mathbf{A}_0\| \cdot \prod_{j=0}^{k}\left(1 + \frac{1}{2}\nu^{2^j}\right) \le \|\mathbf{A}_0\|\left(1 + \frac{1}{2}\sum_{i=1}^{2^{k+1}-1}\nu^i\right) =$$
$$= \|\mathbf{A}_0\|\left(1 + \frac{1}{2}\cdot\frac{\nu - \nu^{2^{k+1}}}{1 - \nu}\right) \le \|\mathbf{A}_0\| \cdot \frac{2 - \nu}{2 - 2\nu} \le \lambda$$

(последнее, в силу наложенного на ν условия (7.69)).

Так как $\left\|F\left(\mathbf{x}^{(0)}\right)\right\| = p_0$, согласно заданию (7.71), а $\|\mathbf{B}_0\| \le b_0$ по условию (7.68) и, кроме того, $\|\mathbf{A}_0\| \le \lambda$ (ибо в противном случае должно быть $\nu < 0$, что противоречило бы определению ν в (7.69)), то в соответствии с принципом математической индукции доказываемое мажорирование действительно имеет место при любом $k \in \mathbf{N}_0$.

Итак, из предыдущего отберем только нужную для применения теоремы 7.3 информацию. Имеем:

$$\left\|\mathbf{x}^{(k+1)}-\mathbf{x}^{(k)}\right\| \le \|\mathbf{A}_k\|\cdot\left\|F\left(\mathbf{x}^{(k)}\right)\right\| \le \lambda p_k,$$

$$\left\|F\left(\mathbf{x}^{(k)}\right)\right\| \le p_k,$$

$$p_{k+1}=G_0 p_k^2, \quad \text{где} \quad G_0=4L\lambda^2, \quad p_0=\left\|F\left(\mathbf{x}^{(0)}\right)\right\|$$

$$(\text{т.е. } \mu=2, \quad \nu=G_0p_0=4L\lambda^2p_0).$$

Поскольку здесь на величину ν наложено более сильное, чем в теореме 7.3, ограничение (7.69), с этим ν заключительная часть теоремы 7.3 будет тем более верна. Теорема доказана.

Условие (7.69) теоремы 7.7 можно трактовать так: параметр $\lambda>0$ должен удовлетворять кубическому неравенству

$$4Lp_0\lambda^3-2L\|\mathbf{A}_0\|p_0\lambda^2-\lambda+\|\mathbf{A}_0\|\le 0.$$

Исследование этого неравенства приводит к ряду следствий [35]; наиболее простым из них (в некотором смысле) является

Следствие 7.1. *Квадратичная сходимость начатого с* $\mathbf{x}^{(0)}$, $\mathbf{A}_0$ *метода* (7.11) *обеспечивается выполнением условий*

$$\left\|\mathbf{E}-F'\left(\mathbf{x}^{(0)}\right)\mathbf{A}_0\right\| \le 0.109, \quad L\|\mathbf{A}_0\|^2\left\|F\left(\mathbf{x}^{(0)}\right)\right\| \le 0.0567,$$

$$S\left(\mathbf{x}^{(0)},\ r:=2.46\cdot\|\mathbf{A}_0\|\cdot\left\|F\left(\mathbf{x}^{(0)}\right)\right\|\right)\subseteq M.^{*)}$$

Привлечение общих теорем 7.3, 7.4 о сходимости итерационных последовательностей позволяет сформулировать и доказать утверждения типа теоремы 7.5 о квадратичной сходимости основного метода Ньютона и для случая, когда итерационная последовательность $\left(\mathbf{x}^{(k)}\right)$ задается полюсным методом Ньютона (7.53). При этом требование 1) теоремы 7.5 остается тем же, а требование 2) о существовании и равномерной ограниченности матрицы, обратной к матрице Якоби $F'(x)$, в n-мерной области M заменяется таким же требованием к «смещенной» матрице Якоби $F'(\mathbf{x})+\mathbf{A}$, где $\mathbf{A}=\mathbf{A}(\mathbf{x})$ — это определенная посредством (7.51) матрица $\mathbf{A}_k=\mathbf{A}\left(\mathbf{x}^{(k)}\right)$, в которой $\mathbf{x}^{(k)}$ заменяются на $\mathbf{x}$.

---

*) Как показывают более тонкие исследования, фигурирующие здесь постоянные 0.109 и 0.0567 сильно занижены (см., например, [203], где в этой роли выступает одна постоянная 0.25).

Замечание 7.7. Приведенный в этом параграфе цикл теорем можно значительно расширить и обобщить.

Во-первых, без каких-либо существенных изменений все эти утверждения могут быть доказаны для нелинейных операторных уравнений в банаховых пространствах.

Во-вторых, вместо условия Липшица на $F'(\mathbf{x})$ (которое, кстати, легко можно заменить требованием ограниченности $\|F''(\mathbf{x})\|$ в M постоянной L) можно вставить более слабое ***условие Гёльдера***

$$\|F'(\mathbf{x})-F'(\tilde{\mathbf{x}})\|\le L\|\mathbf{x}-\tilde{\mathbf{x}}\|^{\alpha}, \quad \alpha\in(0,\ 1],$$

частным случаем которого является условие Липшица. При этом порядок μ метода в теоремах типа теорем 7.5, 7.7 может быть установлен «плавающим» от 1 до $1+\alpha$ в зависимости от жесткости требований, накладываемых на исходные данные [39, 40].

В-третьих, общие теоремы 7.3 и 7.4 позволяют исследовать сходимость методов более высоких порядков как с аппроксимацией обратных к матрицам Якоби матриц, так и без нее, например, методов третьего порядка (7.9) и (7.12).

Использование техники оценочных функций, ключ к которой можно найти в книге [86] Л. Коллатца, а также обобщение леммы 7.2 на случай, когда условие Липшица или Гёльдера накладывается на производные более высоких порядков, позволяют указывать условия сходимости семейства методов вида

$$\mathbf{x}^{(k+1)}=\mathbf{x}^{(k)}-Q\left(\mathbf{x}^{(k)},\ \mathbf{A}_k\right),$$

содержащего изученные здесь методы и ряд других методов.

В заключение отметим, что применение теорем сходимости типа теорем 7.5–7.7 и других подобных утверждений для конкретных нелинейных систем упирается, как правило, в проблему нахождения постоянных Липшица L для производных векторных функций. Даже в простейшем случае, когда L находится из условия $L\ge\|F''(\mathbf{x})\|$, требуется сделать оценку величины нормы функциональной матрицы Гессе $\left(\dfrac{\partial^2 f_m}{\partial x_i\partial y_j}\right)_{i,j,m=1}^{n}$ в некоторой n-мерной области M, что весьма непросто.

УПРАЖНЕНИЯ

7.1. Проведите доказательство теоремы 7.1 по аналогии с доказательством теоремы 6.1.

7.2. Проведите доказательство теоремы 7.2 по аналогии с доказательством теоремы 6.2, используя лемму 7.1.

7.3. Можно ли утверждать, что система

$$\begin{cases} x = 0.1\sin x + 0.3\cos y - 0.4, \\ y = 0.2\cos x - 0.1\sin y - 0.3 \end{cases}$$

имеет и притом единственное вещественное решение? Почему? Сделайте 5 итераций МПИ (7.3), начиная процесс с нулевого начального вектора, и оцените погрешность с помощью априорной оценки теоремы 7.1. Найдите решение с точностью $\varepsilon = 10^{-6}$, останавливая процесс вычислений на основе апостериорной оценки погрешности. Сделайте 5 шагов метода покоординатных итераций (7.4). Сравните результат с предыдущим.

7.4. Дана система

$$\begin{cases} x^3 - y^3 + 0.1 = 0, \\ xy - 0.95 = 0 \end{cases}$$

и точка $(x_0;\ y_0) = (1;1)$. Запишите для этой системы основной (7.6) и модифицированный (7.8) методы Ньютона и сделайте по 2–3 итерационных шага. Попытайтесь применить здесь какие-нибудь подходящие теоремы сходимости (см. теоремы 7.1, 7.2, 7.5, 7.6).

7.5. Сравните по числу арифметических операций, приходящихся на реализацию одного итерационного шага при решении n-мерной нелинейной системы, следующие методы:

а) метод Ньютона в явной форме (7.6);
б) метод Ньютона в неявной форме (7.7);
в) аппроксимационный аналог метода Ньютона (7.11);
г) простейший метод секущих (7.14);
д) разностный метод Ньютона (7.13);
е) метод секущих Бройдена (7.32)–(7.34) или (7.26), (7.35);
ж) полюсный метод Ньютона (7.50)–(7.52) или (7.53).

Подсчет вычислительных затрат вести, предполагая, что решение линейных систем в (7.7), (7.32), (7.50) и обращение матриц в (7.6), (7.14), (7.13), (7.26) и (7.53) производится методом Гаусса и что вычисление значений функций и производных при этом во внимание не принимается, т.е. эти значения считаются уже найденными.

7.6. Запишите алгоритм решения нелинейных систем такой модификацией метода Ньютона, при которой частные производные в матрицах Якоби аппроксимируются симметричными разностными отношениями:

$$\frac{\partial f_i}{\partial x_j} \approx \frac{f_i(x_1, \ldots, x_j + h_j, \ldots, x_n) - f_i(x_1, \ldots, x_j - h_j, \ldots, x_n)}{2h_j}.$$

Увеличатся ли вычислительные затраты при переходе от (7.13) к такому симметричному разностному методу?

7.7. Дополните рассмотрение численного примера § 7.7 (табл. 7.1) применением методов третьего порядка (7.9) (или, что то же, (7.10)) и (7.12). Проведите сравнительный анализ полученных результатов.

7.8. Запишите аппроксимационные аналоги (типа ААМН (7.11)) для методов секущих (7.14) и (7.26). Протестируйте новые методы на системе (7.58).

7.9. Выведите расчетные формулы метода Брауна (см. § 7.3) для трехмерного случая. Опробуйте полученные формулы на системе

$$\begin{cases} x^2 + y^2 - z = 0, \\ x^2 + y^2 - z^2 = 0, \\ \ln x - \sqrt{y} + 0.8 = 0, \end{cases}$$

взяв начальное приближение $x_0 = 0.5$, $y_0 = 0.5$, $z_0 = 0.5$. Проверьте результаты тем же методом Брауна, но примененным к двумерной системе, к которой легко перейти от данной исключением z.

7.10. Убедитесь, что n-полюсный метод Ньютона (7.53) действительно обобщает однополюсный (5.36) и двухполюсный (7.44) методы Ньютона при $n = 1$ и $n = 2$ соответственно. При каких значениях параметров n-полюсный метод превращается в основной метод Ньютона (7.6)?

7.11. Составьте гибридный алгоритм, осуществляющий поиск решения нелинейной системы сначала методом градиентного спуска, а затем методом Ньютона или какой-либо его быстросходящейся модификацией. Примените построенный алгоритм к системе из предыдущего упражнения при различных начальных точках (в частности, при $x_0 = y_0 = z_0 = 0$).

ГЛАВА 8 ПОЛИНОМИАЛЬНАЯ ИНТЕРПОЛЯЦИЯ

Прежде всего, обсуждаются различные постановки задачи аппроксимации функций одной переменной, в частности, полиномиальной аппроксимации. Далее конкретизируется задача интерполяции, определяется решающий эту задачу интерполяционный многочлен Лагранжа, доказывается его единственность, выводится остаточный член. Рассматривается итерационный принцип вычисления промежуточных значений таблично заданных (сеточных) функций с помощью лагранжевой интерполяции, известный как интерполяционная схема Эйткена. Для случая равноотстоящих узлов вводятся конечные разности, отмечаются их простейшие свойства, строятся конечно-разностные интерполяционные многочлены, в структуру которых заложено убывание значимости каждого последующего слагаемого. Формулы такой же структуры, использующие аппарат разделенных разностей, выводятся и для неравных промежутков между узлами. Завершается глава изучением задачи обратного интерполирования и задачи интерполирования с кратными узлами; последняя объединяет в себе лагранжеву интерполяцию и локальную аппроксимацию функций по формуле Тейлора.

8.1. ЗАДАЧА И СПОСОБЫ АППРОКСИМАЦИИ ФУНКЦИЙ

В основе большинства численных методов математического анализа лежит подмена одной функции $f(x)$ (известной, неизвестной или частично известной) другой функцией $\varphi(x)$, близкой к $f(x)$ и обладающей «хорошими» свойствами, позволяющими легко производить над нею те или иные аналитические или вычислительные операции. Будем называть такую подмену ***аппроксимацией****) или просто ***приближением*** функции $f(x)$ функцией $\varphi(x)$. Для того, чтобы построить какую-то разумную теорию таких приближений и предложить конкретные способы получения аппроксимирующих функций $\varphi(x)$ по заданным тем

*) Часто термин «аппроксимация функций» используется в более узком смысле, чем это принимается здесь.

или иным образом аппроксимируемым функциям $f(x)$, предварительно следует ответить на ряд вопросов.

1) *Что известно о функции* $f(x)$? Задана ли она своим аналитическим выражением или таблицей своих значений, какова степень ее гладкости и доступны ли значения ее производных, как расположены точки в интересующей части области определения $f(x)$, где известны ее значения, и можно ли их задавать по своему усмотрению, и т.п.

2) *Какому классу (семейству) функций должна принадлежать функция* $\varphi(x)$? Какие дополнительные требования предъявляются к $\varphi(x)$, выделяющие ее из заданного класса?

3) *Что понимать под близостью между* $f(x)$ *и* $\varphi(x)$; *иначе, какой принять* ***критерий согласия*** *между ними*? Говоря языком функционального анализа, по метрике какого пространства должно быть малым расстояние между $f(x)$ и $\varphi(x)$?

Как видим, задача аппроксимации функции $f(x)$ функцией $\varphi(x)$ состоит в построении для заданной функции $f(x)$ такой функции $\varphi(x)$, что

$$f(x)\approx\varphi(x), \tag{8.1}$$

причем левая часть приближенного равенства (8.1) должна быть обусловлена ответами на вопросы первой группы, правая часть — второй группы, а ответ на вопрос 3) должен уточнить значение связывающего $f(x)$ и $\varphi(x)$ символа «≈».

Прежде всего, определимся с ответом на второй вопрос. *Договоримся использовать в качестве аппроксимирующих функций* $\varphi(x)$ *только многочлены или функции, составленные из многочленов**)*; в таком случае будем говорить о ***полиномиальной аппроксимации*** или ***кусочно-полиномиальной аппроксимации*** соответственно.

По сравнению с другими семействами функций, пригодных для построения теории приближений, например, таких, как тригонометрические или показательные функции, рациональные функции или всплески [18], для вычислительной математики многочлены привлекательны тем, что они являются линейными функциями своих параметров (коэффициентов), и их вычисление сводится к выполнению конечного числа простейших арифметических операций — сложения и умножения.

Будем считать, что аппроксимация функции $f(x)$ произво-

*) За небольшим исключением в § 10.1.

дится с помощью многочленов степени $n \in \mathbf{N}_0$. Тогда в зависимости от выбора критерия согласия и, в частности, от количества *точек согласования* $f(x)$ с $\varphi(x)$ (будем называть их ***узлами***), т.е. точек, в которых известна информация об $f(x)$ и, возможно, ее производных, можно рассмотреть разные конкретные способы аппроксимации. Некоторые из них, нашедшие отражение в табл.8.1, будут изучаться ниже. А именно, в данной главе будет достаточно подробно рассматриваться классическая лагранжева интерполяция (клетка 1.2 в табл.8.1), служащая основой многих численных методов, в частности, приближенного дифференцирования и интегрирования. Здесь же можно будет получить представление об интерполяционном многочлене Эрмита (клетка 2.1), частными случаями которого являются, с одной стороны, многочлен Тейлора (при $i = 1$), с другой — многочлен Лагранжа (при $i = n + 1$). Понятие о многочленах наилучших равномерных приближений (клетка 3.3) дается в главе 9, наилучшие среднеквадратические приближения (клетка 4.3) освещаются в главе 10, а кусочно-полиномиальной аппроксимации (клетка 1.3) и сплайнам (клетка 2.3) посвящена глава 11.

Таблица 8.1

Тип функции $\varphi(x)$, аппроксимирующей $f(x)$, выраженной через многочлены степени n (при разных условиях согласования)

Критерий согласия	Количество точек согласования (узлов)		
	$(1 \le)\ i\ (\le n+1)$	$n+1$	$m\ (\ge n+1)$
Совпадение значений функций $f(x)$ и $\varphi(x)$ в узлах	1.1	1.2 Интерполяционный многочлен Лагранжа	1.3 Кусочно-полиномиальная функция
Совпадение в узлах функций $f(x)$ и $\varphi(x)$ и некоторых их производных	2.1 Интерполяционный многочлен Эрмита	2.2	2.3 Интерполяционный сплайн
Минимум максимального отклонения (на отрезке)	3.1	3.2	3.3 Многочлен наилучшего равномерного приближения
Минимум среднеквадратического отклонения (в узлах или на отрезке)	4.1	4.2	4.3 Многочлен наилучшего среднеквадратического приближения

8.2. ИНТЕРПОЛЯЦИОННЫЙ МНОГОЧЛЕН ЛАГРАНЖА

Пусть в точках $x_0, x_1, \ldots, x_n$ таких, что $a \le x_0 < \ldots < x_n \le b$, известны значения функции $y = f(x)$, т.е. на отрезке $[a, b]$ задана ***табличная (сеточная) функция***

$$f(x):\quad \begin{array}{c|c|c|c|c} x & x_0 & x_1 & \ldots & x_n \\ \hline y & y_0 & y_1 & \ldots & y_n \end{array}. \tag{8.2}$$

Функция $\varphi(x)$ называется ***интерполирующей*** или ***интерполяционной*** для $f(x)$ на $[a, b]$, если ее значения $\varphi(x_0), \varphi(x_1), \ldots, \varphi(x_n)$ в заданных точках $x_0, x_1, \ldots, x_n$, называемых ***узлами интерполяции***, совпадают с заданными значениями функции $f(x)$, т.е. с $y_0, y_1, \ldots, y_n$ соответственно*). Геометрически факт интерполирования означает, что график функции $\varphi(x)$ проходит так, что, по меньшей мере, в $n + 1$ заданных точках он пересекает или касается графика функции $f(x)$ (рис.8.1).

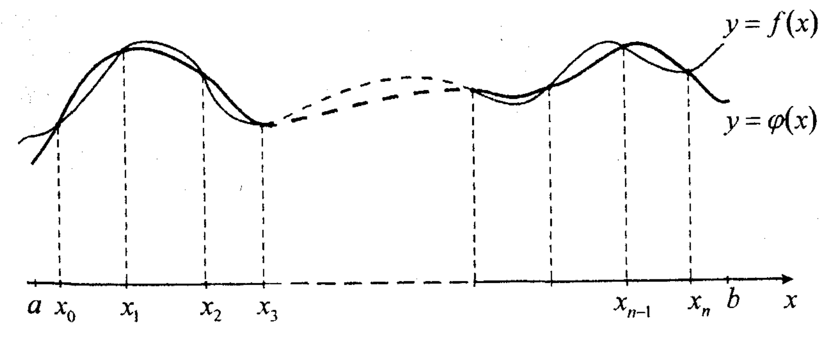


Рис.8.1. Геометрическая интерпретация задачи интерполирования

Легко представить, что таких графиков $\varphi(x)$, проходящих через заданные точки, можно изобразить сколько угодно, и они могут отличаться от графика $f(x)$ сколь угодно сильно, если не накладывать на $\varphi(x)$ и $f(x)$ определенных ограничений.

---

*) Латинское слово *interpolatio* переводится как обновление, изменение, переделка. Обычно термин ***интерполяция*** (или иначе, ***интерполирование***) означает п р о ц е с с построения интерполяционной функции или процесс нахождения промежуточных значений табличной функции. Этот термин введен в 1656 году английским математиком Джоном Валлисом (Уоллисом, 1616–1703 гг.); кстати отметим, что годом раньше он ввел ныне общепринятый символ бесконечности (∞) [199].

В соответствии с договоренностью предыдущего параграфа будем считать, что интерполяционная функция $\varphi(x)$ есть многочлен степени n. Тогда **задача интерполяции**, точнее, полиномиальной, алгебраической или параболической интерполяции (поскольку график любого многочлена называют параболой) формулируется так:

для функции $f(x)$, заданной таблицей (8.2), *найти многочлен $P_n(x)$ такой, что выполняется совокупность* ***условий интерполяции***

$$P_n(x_i) = y_i \quad \forall\, i \in \{0, 1, \dots, n\}. \tag{8.3}$$

Найти многочлен $P_n(x)$ — это значит, учитывая его каноническую форму

$$P_n(x) = a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n, \tag{8.4}$$

найти его $n+1$ коэффициент $a_0, a_1, \dots, a_n$. Для этого имеется как раз $n+1$ условие (8.3). Таким образом, чтобы многочлен (8.4) был интерполяционным для функции (8.2), нужно, чтобы его коэффициенты $a_0, a_1, \dots, a_n$ удовлетворяли системе уравнений

$$\begin{cases} a_0 + a_1 x_0 + a_2 x_0^2 + \dots + a_n x_0^n = y_0, \\ a_0 + a_1 x_1 + a_2 x_1^2 + \dots + a_n x_1^n = y_1, \\ \dots\dots\dots\dots\dots\dots\dots \\ a_0 + a_1 x_n + a_2 x_n^2 + \dots + a_n x_n^n = y_n. \end{cases}$$

Из курса алгебры известно, что определитель этой линейной системы (так называемый определитель Вандермонда) отличен от нуля, т.е. решение этой системы существует и единственно. Однако практическое построение интерполяционного многочлена таким путем малоэффективно. Поэтому изберем другой путь.

Будем строить многочлен n-й степени $L_n(x)$ в виде линейной комбинации $\sum_{i=0}^{n} c_i l_i(x)$ многочленов n-й же степени $l_i(x)$ $(i = 0, 1, \dots, n)$ *). Для того, чтобы такой многочлен был интерполяционным для функции $f(x)$, достаточно зафиксировать в качестве коэффициентов c_i этой линейной комбинации заданные в табл.8.2 значения $y_i = f(x_i)$, а от базисных многочленов $l_i(x)$

*) В отличие от общепринятых для многочленов обозначений, здесь индекс у $l_i(x)$ показывает не степень, а номер многочлена от 0 до n.

потребовать выполнения условия

$$l_i(x_j) = \delta_{ij} := \begin{cases} 0, \text{ если } j \ne i \\ 1, \text{ если } j = i \end{cases} \quad \forall\, j, i \in \{0, 1, \dots, n\}. \tag{8.5}$$

В таком случае для многочлена

$$L_n(x) = \sum_{i=0}^{n} y_i l_i(x)$$

в каждом узле x_j $(j \in \{0, 1, \dots, n\})$, в силу (8.5), справедливо

$$L_n(x_j) = l_0(x_j) y_0 + \dots + l_{j-1}(x_j) y_{j-1} + l_j(x_j) y_j + l_{j+1}(x_j) y_{j+1} + \dots$$
$$\dots + l_n(x_j) y_n = 0 + \dots + 0 + y_j + 0 + \dots + 0 = y_j,$$

т.е. выполняются условия интерполяции (8.3).

Чтобы конкретизировать базисные многочлены $l_i(x)$, учтем, что они должны удовлетворять условиям (8.5). Равенство нулю i-го многочлена во всех узлах, кроме i-го, означает, что $l_i(x)$ можно записать в виде

$$l_i(x) = A_i (x - x_0) \dots (x - x_{i-1})(x - x_{i+1}) \dots (x - x_n),$$

а коэффициент A_i этого представления легко получается из содержащегося в (8.5) требования $l_i(x_i) = 1$. Подставляя в выражение $l_i(x)$ значение $x = x_i$ и приравнивая результат единице, получаем

$$A_i = \frac{1}{(x_i - x_0) \dots (x_i - x_{i-1})(x_i - x_{i+1}) \dots (x_i - x_n)}.$$

Таким образом, ***базисные многочлены Лагранжа*** имеют вид

$$l_i(x) = \frac{(x - x_0)(x - x_1) \dots (x - x_{i-1})(x - x_{i+1}) \dots (x - x_n)}{(x_i - x_0)(x_i - x_1) \dots (x_i - x_{i-1})(x_i - x_{i+1}) \dots (x_i - x_n)},$$

а искомый ***интерполяционный многочлен Лагранжа*** есть

$$L_n(x) = \sum_{i=0}^{n} \frac{(x - x_0) \dots (x - x_{i-1})(x - x_{i+1}) \dots (x - x_n)}{(x_i - x_0) \dots (x_i - x_{i-1})(x_i - x_{i+1}) \dots (x_i - x_n)} y_i. \tag{8.6}$$

Заметим, что *числитель, фигурирующий в записи i-го слагаемого $L_n(x)$ дроби, представляет собой произведение разностей между переменной x и всеми узлами, кроме i-го, а знаменатель — произведение разностей между i-м узлом и всеми остальными.*

В качестве примера запишем интерполяционные многочлены Лагранжа первой и второй степени.

При $n=1$ информация об интерполируемой функции $y=f(x)$ сосредоточена в двух точках: $(x_0;y_0)$ и $(x_1;y_1)$. Многочлен Лагранжа в этом случае составляется с помощью двух базисных многочленов первой степени ($l_0(x)$ и $l_1(x)$) и имеет вид

$$L_1(x) = \frac{x-x_1}{x_0-x_1}y_0 + \frac{x-x_0}{x_1-x_0}y_1. \tag{8.7}$$

При $n=2$ по трехточечной таблице

$$f(x):\quad \begin{array}{c||c|c|c} x & x_0 & x_1 & x_2 \\ \hline y & y_0 & y_1 & y_2 \end{array}$$

можно образовать три базисных многочлена ($l_0(x)$, $l_1(x)$ и $l_2(x)$) и, соответственно, интерполяционный многочлен Лагранжа второй степени

$$L_2(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}y_0 + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}y_1 + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}y_2. \tag{8.8}$$

Приближенные равенства

$$f(x)\approx L_1(x) \quad \text{и} \quad f(x)\approx L_2(x)$$

называют соответственно ***формулами линейной и квадратичной интерполяции***. Геометрически они означают подмену графика функции $y=f(x)$ на некотором отрезке $[a,b]$ оси абсцисс, содержащем точки x_0, x_1 в первом и x_0, x_1, x_2 во втором случаях, соответствующими участками прямой линии и квадратичной параболы, проходящих через заданные точки координатной плоскости.

Такая простейшая интерполяция широко применялась при составлении различных таблиц значений функций для их пополнения промежуточными значениями (что и являлось основной задачей интерполяции на ранней стадии развития вычислительной математики). В связи с этим, заметим, что иногда термину *интерполяция* противопоставляется термин ***экстраполяция***. В таких случаях речь идет о том, что под интерполяцией понимается нахождение промежуточных значений таблично заданной функции строго в н у т р и таблицы, тогда как экстраполяция*) предполагает использование интерполяционного многочлена, построенного по значениям функции $f(x)$ в точках $x_0, x_1, \ldots, x_n$, для нахождения ее приближенных значений з а п р е д е л а м и промежутка $[x_0, x_n]$.

Вернемся к изучению интерполяционного многочлена Лагранжа (8.6).

Покажем его **единственность** (от противного). Предположим, что наряду с $L_n(x)$ имеется другой многочлен n-й степени $Q_n(x)$, решающий ту же задачу интерполяции, т.е. удовлетворяющий условиям интерполяции типа (8.3):

$$Q_n(x_i) = y_i \quad \forall\ i \in \{0, 1, \ldots, n\}.$$

Образуем новый многочлен как разность между $L_n(x)$ и $Q_n(x)$. Этот многочлен $P_n(x) := L_n(x) - Q_n(x)$ имеет степень не выше n и во всех $n+1$ узлах $x_0, x_1, \ldots, x_n$ обращается в нуль, в силу равенства значений $Q_n(x_i)$ и $L_n(x_i)$ одним и тем же числам y_i. Получается, что точки $x_0, x_1, \ldots, x_n$ служат корнями многочлена $P_n(x)$. Но по следствию из основной теоремы алгебры многочленов $P_n(x)$ не может иметь более n корней. Полученное противоречие означает, что многочлены $Q_n(x)$ и $L_n(x)$ должны полностью совпадать, т.е. *по заданным $n+1$ значениям функции можно построить единственный интерполяционный многочлен.*

Пусть для данной функции $f(x)$ интерполяционный многочлен $L_n(x)$ построен, т.е. для приближенного представления функции $f(x)$ на отрезке $[a,b]\supseteq[x_0,x_n]$ применяется ***интерполяционная формула***

$$f(x)\approx L_n(x). \tag{8.9}$$

Естественно встает вопрос: какова погрешность такого приближенного равенства? Иначе, сколь велико может быть различие между значениями интерполируемой функции $f(x)$ и соответствующими значениями интерполяционного многочлена Лагранжа $L_n(x)$ в точках отрезка $[a,b]$, не совпадающих с узловыми точками?

---

*) Указанный здесь узкий смысл терминов «интерполирование» и «экстраполирование» становится очевидным, если учитывать их латинское происхождение: «inter» и «extra» означают соответственно «между» и «вне», а «polire» — «делать гладким» [200]; таким образом, интерполирование — это сглаживание между узлами, а экстраполирование — сглаживание вне таблицы.

Для совершенно произвольной функции $f(x)$ такая постановка вопроса о погрешности интерполяции заведомо некорректна, поскольку функций, для которых построенный (единственный!) многочлен $L_n(x)$ будет интерполяционным, бесконечно много; легко представить себе эту ситуацию графически: через $n+1$ заданных точек с координатами $(x_0;\ y_0)$, $(x_1;\ y_1)$, ..., $(x_n;\ y_n)$, согласно доказанному, можно провести единственную параболу — график многочлена степени n — и, в то же время, можно изобразить сколько угодно графиков других функций, как угодно сильно отличающихся от этой параболы. Этот факт говорит о том, что заключить величину этого отклонения, т.е. погрешность интерполяции, в определенные рамки невозможно, если не наложить каких-то ограничений на гладкость интерполируемой функции $f(x)$ и на расположение узлов интерполяции $x_0, x_1, \ldots, x_n$ на отрезке $[a, b]$.

Будем выяснять величину отклонения $f(x)$ от $L_n(x)$ в произвольной точке $x \in [a, b]$, иначе, величину ***остаточного члена***

$$R_n(x) := f(x) - L_n(x)$$

интерполяционной формулы Лагранжа (8.9) в предположении, что $f(x) \in C^{n+1}_{[a,\ b]}$, т.е. данная функция $n+1$ раз непрерывно дифференцируема.

Обозначим

$$\Pi_{n+1}(x) := \prod_{i=0}^{n}(x - x_i) = (x - x_0)(x - x_1)\ldots(x - x_n) \tag{8.10}$$

— определенный через узлы $x_0, x_1, \ldots, x_n$ многочлен $(n+1)$-й степени. Через него введем в рассмотрение функцию

$$u(x) := f(x) - L_n(x) - c\Pi_{n+1}(x), \tag{8.11}$$

где c — некоторая постоянная (параметр).

Так как в точках $x = x_0, x_1, \ldots, x_n$ многочлен $\Pi_{n+1}(x)$ обращается в нуль, согласно его конструкции, а $f(x) - L_n(x) = 0$ в этих точках по условиям интерполяции, то и $u(x_i) = 0$ при $i = 0, 1, \ldots, n$, т.е. функция $u(x)$ имеет на отрезке $[a, b]$ по меньшей мере $n+1$ корень. Подберем параметр c так, чтобы $u(x)$ имела заведомо еще и $(n+2)$-й корень в какой-то фиксированной точке $\bar{x}$ $(\neq x_i)$ промежутка $[a, b]$. Имеем:

$$u(\bar{x}) = 0 \quad \Leftrightarrow \quad f(\bar{x}) - L_n(\bar{x}) = c\Pi_{n+1}(\bar{x}) \quad \Leftrightarrow \quad c = \frac{f(\bar{x}) - L_n(\bar{x})}{\Pi_{n+1}(\bar{x})},$$

причем такое значение c обязательно найдется, поскольку $\Pi_{n+1}(x) = 0$ только в узлах x_i.

Пусть для определенности $\bar{x} \in (x_i,\ x_{i+1})$. Тогда можно утверждать, что при найденном c функция $u(x)$ равна нулю на концах $n+1$ отрезков $[x_0,\ x_1]$, $[x_1,\ x_2]$, ..., $[x_i,\ \bar{x}]$, $[\bar{x},\ x_{i+1}]$, ..., $[x_{n-1},\ x_n]$. Значит, к функции $u(x)$ на каждом из этих отрезков применима теорема Ролля, т.е. внутри каждого из этих отрезков существует, по крайней мере, по одной такой точке, в которой производная функции $u(x)$ обращается в нуль. Эти $n+1$ точки образуют систему из n отрезков, на концах каждого из которых уже функция $u'(x)$ равна нулю, т.е. теперь к производной можно применить теорему Ролля, по которой существует n нулей второй производной функции $u(x)$. Продолжая процесс таких рассуждений далее, в конце концов, приходим к выводу о существовании такой точки $\xi \in (x_0,\ x_n) \subseteq (a,\ b)$, что $u^{(n+1)}(\xi) = 0$. Учитывая, что n-я производная многочлена n-й степени постоянна, а $(n+1)$-я равна нулю, находим выражение $(n+1)$-й производной функции $u(x)$, заданной равенством (8.11):

$$u^{(n+1)}(x) = f^{(n+1)}(x) - 0 - c(n+1)!\ .$$

Итак, существует точка $\xi \in (x_0,\ x_n)$ такая, что

$$f^{(n+1)}(\xi) - c(n+1)! = 0, \qquad \text{т.е.} \quad c = \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Это значение c должно совпадать с выбранным ранее, т.е. должно выполняться равенство

$$\frac{f(\bar{x}) - L_n(\bar{x})}{\Pi_{n+1}(\bar{x})} = \frac{f^{(n+1)}(\xi)}{(n+1)!},$$

откуда получаем

$$f(\bar{x}) - L_n(\bar{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\Pi_{n+1}(\bar{x}).$$

Так как в качестве $\bar{x}$ могла быть взята любая точка x из промежутка $[a, b]$, не совпадающая ни с какой узловой, расфиксируем (или, как еще говорят, разморозим) точку $\bar{x}$, т.е. заменим ее в последнем равенстве произвольной точкой $x \neq x_i$, в результате чего приходим к выражению остаточного члена

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\Pi_{n+1}(x). \tag{8.12}$$

Знание остаточного члена в предположении $(n+1)$-кратной дифференцируемости $f(x)$ позволяет записать точное представление $f(x)$ через ее интерполяционный многочлен $L_n(x)$:

$$f(x) = L_n(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!}\Pi_{n+1}(x), \tag{8.13}$$

где ξ — некоторая (вообще говоря, неизвестная, причем зависящая от x) точка из промежутка интерполяции (a, b), а $\Pi_{n+1}(x)$ — определенный в (8.10) многочлен *).

Теперь можно ставить и пытаться отвечать на вопросы о погрешности приближенного вычисления значения $f(x)$ с помощью $L_n(x)$ в какой-либо конкретной точке промежутка $[a, b]$, о величине максимальной погрешности, допускаемой при подмене функции $f(x)$ многочленом $L_n(x)$ на этом промежутке, о сходимости интерполяционного процесса, т.е. о том, имеет ли место $\rho(f(x), L_n(x)) \xrightarrow[n\to\infty]{} 0$ по метрике $\rho(\cdot,\cdot)$ того или иного определенного на $[a, b]$ функционального пространства.

Так, если известна величина

$$M_{n+1} := \max_{x\in[a,\, b]} \left|f^{(n+1)}(x)\right|,$$

то оценить абсолютную погрешность интерполяционной формулы (8.9) в любой точке $\widetilde{x} \in [a, b]$ можно с помощью неравенства

$$|R_n(\widetilde{x})| = |f(\widetilde{x}) - L_n(\widetilde{x})| \le \frac{M_{n+1}}{(n+1)!}|\Pi_{n+1}(\widetilde{x})|. \tag{8.14}$$

Максимальная погрешность интерполирования на отрезке $[a, b]$ оценивается величиной

$$\max_{x\in[a,\, b]} |R_n(x)| \le \frac{M_{n+1}}{(n+1)!}\max_{x\in[a,\, b]} |\Pi_{n+1}(x)|. \tag{8.15}$$

Так как максимумы функций $f^{(n+1)}(x)$ и $\Pi_{n+1}(x)$ достигаются, вообще говоря, в разных точках отрезка $[a, b]$, то более точной, но более трудно реализуемой по сравнению с (8.15) следует считать оценку

$$\max_{x\in[a,b]} |R_n(x)| \le \frac{1}{(n+1)!}\max_{x\in[a,b]} \left|f^{(n+1)}(x)\Pi_{n+1}(x)\right|.$$

---

*) Формула (8.12) устанавливалась для значений $x \ne x_i$. Но при $x = x_i$ левая и правая части (8.12) равны нулю, следовательно, формула (8.12), а с нею и представление (8.13), справедливы при любых $x \in [a, b]$.

Пример 8.1. Рассмотрим квадратичную интерполяцию функции $y = \sin x$ на отрезке $\left[0, \frac{\pi}{2}\right]$ по ее трем значениям: $\sin 0 = 0$, $\sin\frac{\pi}{4} = \frac{\sqrt{2}}{2}$, $\sin\frac{\pi}{2} = 1$.

По формуле (8.8) строим многочлен Лагранжа второй степени

$$L_2(x) = \frac{\left(x-\frac{\pi}{4}\right)\left(x-\frac{\pi}{2}\right)}{\left(0-\frac{\pi}{4}\right)\left(0-\frac{\pi}{2}\right)}\cdot 0 + \frac{(x-0)\left(x-\frac{\pi}{2}\right)}{\left(\frac{\pi}{4}-0\right)\left(\frac{\pi}{4}-\frac{\pi}{2}\right)}\cdot\frac{\sqrt{2}}{2} + \frac{(x-0)\left(x-\frac{\pi}{4}\right)}{\left(\frac{\pi}{2}-0\right)\left(\frac{\pi}{2}-\frac{\pi}{4}\right)}\cdot 1,$$

или после преобразований

$$L_2(x) = \frac{8}{\pi^2}x\left[\left(1-\sqrt{2}\right)x + \left(\frac{\sqrt{2}}{2}-\frac{1}{4}\right)\pi\right]. \tag{8.16}$$

Остаточный член для этого случая получаем по формуле (8.12), учитывая, что $n = 2$, $(\sin x)''' = -\cos x$ и $\Pi_3(x) = x\left(x-\frac{\pi}{4}\right)\left(x-\frac{\pi}{2}\right)$.

Имеем:

$$R_2(x) = \sin x - L_2(x) = \frac{-\cos\xi}{3!}x\left(x-\frac{\pi}{4}\right)\left(x-\frac{\pi}{2}\right).$$

Так как точка $\xi \in \left(0, \frac{\pi}{2}\right)$ неизвестна, можно делать лишь оценки $|R_2(x)|$, полагая $M_3 = \max_{x\in\left[0,\frac{\pi}{2}\right]} |-\cos x| = 1$. Найдя максимальное значение $|\Pi_3(x)|$, реализуемое в двух точках данного отрезка $x_{1,2} = \frac{\pi}{12}\left(3 \pm \sqrt{3}\right)$ и не превосходящее 0.568, по формуле (8.15) оцениваем сверху величину допустимого отклонения дуги параболы (8.16) от данной синусоиды на промежутке интерполирования:

$$\max_{x\in\left[0,\frac{\pi}{2}\right]} |\sin x - L_2(x)| \le \frac{1}{3!}\max_{x\in\left[0,\frac{\pi}{2}\right]} |L_2(x)| < \frac{0.568}{6} \approx 0.095.$$

Подставим в полученный интерполяционный многочлен (8.16) контрольную точку $\widetilde{x} = \frac{\pi}{6}$. Получим приближенное значение $L_2\left(\frac{\pi}{6}\right) = \frac{4\sqrt{2}-1}{9} \approx 0.517$, отличающееся от значения $\sin\frac{\pi}{6} = 0.5$ на вели-

чину ≈ 0.017, меньшую, чем это допускается оценкой по формуле (8.14):

$$\left|\sin\frac{\pi}{6} - L_2\left(\frac{\pi}{6}\right)\right| \le \frac{M_3}{3!}\left|\Pi_3\left(\frac{\pi}{6}\right)\right| = \frac{1}{6}\left(\frac{\pi}{6}\right)^3 \approx 0.075.$$

Наблюдаем типичную картину: фактическая погрешность меньше оценки погрешности в точке, которая, в свою очередь, меньше максимальной погрешности на отрезке (т.е. по чебышевской норме).

В заключение этого параграфа заметим, что через введенный в (8.10) многочлен $\Pi_{n+1}(x)$ интерполяционный многочлен Лагранжа (8.6) можно записать в более компактной форме. Для этого достаточно увидеть, что знаменатель фигурирующей там дроби представляет собой значение производной многочлена $\Pi_{n+1}(x)$ в i-м узле, а числитель есть просто $\Pi_{n+1}(x)$ без множителя $x - x_i$.

Таким образом,

$$L_n(x) = \sum_{i=0}^{n} \frac{\Pi_{n+1}(x) y_i}{(x - x_i)\Pi'_{n+1}(x_i)} = \Pi_{n+1}(x)\sum_{i=0}^{n} \frac{y_i}{(x - x_i)\Pi'_{n+1}(x_i)}. \quad (8.6a)$$

Вопросы сходимости $L_n(x)$ к $f(x)$ при $n \to \infty$ будут обсуждаться несколько позже (см. § 8.5).

8.3. ИНТЕРПОЛЯЦИОННАЯ СХЕМА ЭЙТКЕНА

Пусть функция $f(x)$ и расположение узлов $x_0, x_1, \ldots, x_n$ на промежутке интерполяции $[a, b]$ таковы, что имеет место сходимость процесса интерполяции, т.е. $R_n(x) \to 0$ при $n \to \infty$, и пусть решается частная задача вычисления отдельных приближенных значений функции $f(x)$ с помощью вычисления соответствующих им значений интерполяционного многочлена Лагранжа $L_n(x)$. Для построения эффективного способа решения такой частной задачи интерполяции примем во внимание следующие три обстоятельства.

Во-первых, непосредственное использование многочлена Лагранжа в форме (8.6) неудобно из-за его громоздкости (что чревато большими вычислительными затратами). Во-вторых, как правило, заранее неизвестно, какой степени многочлен нужно использовать для интерполирования данной функции с требуемой точностью, а постепенное наращивание точности за счет повторных вычислений значений $L_n(x)$ со все большими показателями степени n (подобно тому, как это можно делать, например, при вычислении значений функции по формуле Тейлора) при прямом применении формулы (8.6) малоприемлемо, в силу плохой перестраиваемости $L_{n-1}(x)$ в $L_n(x)$. В-третьих, при реальном счете всегда следует помнить, что функция $f(x)$ задается таблицей своих п р и б л и ж е н н ы х значений, и каноническое развитие процесса сходимости $L_n(x)$ к $f(x)$ при больших значениях n будет нарушено все возрастающим влиянием на результат исходных ошибок (более подробно об этом см. далее в §§ 8.4, 8.5).

Построим вычислительную схему для получения приближенного значения сеточной функции $f(x)$ в заданной точке $x = \tilde{x}$, в основу которой будет положена лагранжева интерполяция на сетке узлов $x_0, x_1, \ldots, x_n$ и организация вычислений по которой будет иметь итерационный характер. Каждый итерационный шаг в этой схеме заключается в вычислении некоторого определителя второго порядка.

Пусть даны две точки на кривой $y = f(x)$: $(x_0; y_0)$ и $(x_1; y_1)$. Введем функцию $P_{0,1}(x)$ через определитель следующим образом:

$$P_{0,1}(x) := \frac{1}{x_1 - x_0}\begin{vmatrix} x - x_0 & y_0 \\ x - x_1 & y_1 \end{vmatrix}.$$

Раскрыв этот определитель, видим, что

$$P_{0,1}(x) = \frac{x - x_1}{x_0 - x_1} y_0 + \frac{x - x_0}{x_1 - x_0} y_1 = L_1(x),$$

т.е. $P_{0,1}(x)$ совпадает с интерполяционным многочленом Лагранжа первой степени, построенным по данным двум точкам (сравните с (8.7)).

Если взять на кривой $y = f(x)$ точки $(x_1; y_1)$ и $(x_2; y_2)$, то, очевидно, функция

$$P_{1,2}(x) := \frac{1}{x_2 - x_1}\begin{vmatrix} x - x_1 & y_1 \\ x - x_2 & y_2 \end{vmatrix},$$

в развернутой форме имеющая вид

$$P_{1,2}(x) = \frac{x - x_2}{x_1 - x_2} y_1 + \frac{x - x_1}{x_2 - x_1} y_2,$$

тоже является многочленом Лагранжа первой степени, интерполирующим $f(x)$ по точкам $(x_1; y_1)$ и $(x_2; y_2)$.

Считая, что на кривой $y = f(x)$ заданы три точки $(x_0; y_0)$, $(x_1; y_1)$ и $(x_2; y_2)$, с помощью введенных линейных функций $P_{0,1}(x)$ и $P_{1,2}(x)$ образуем новую функцию

$$P_{0,1,2}(x) := \frac{1}{x_2 - x_0}\begin{vmatrix} x - x_0 & P_{0,1}(x) \\ x - x_2 & P_{1,2}(x) \end{vmatrix}.$$

Легко видеть, что эта функция есть многочлен второй степени (точнее, не выше второй). Учитывая, что

$$P_{0,1}(x_0) = P_{1,2}(x_1) = y_0, \quad P_{0,1}(x_1) = y_1 \quad \text{и} \quad P_{1,2}(x_2) = y_2,$$

подстановкой в $P_{0,1,2}(x)$ поочередно значений $x = x_0, x_1, x_2$ убеждаемся, что $P_{0,1,2}(x_0) = y_0$, $P_{0,1,2}(x_1) = y_1$, $P_{0,1,2}(x_2) = y_2$. Таким образом, функция $P_{0,1,2}(x)$ есть многочлен второй степени, решающий задачу параболической интерполяции по трем точкам $(x_0; y_0)$, $(x_1; y_1)$, $(x_2; y_2)$. Но такой многочлен, как доказано в предыдущем параграфе, единствен, следовательно, $P_{0,1,2}(x) = L_2(x)$, где $L_2(x)$ — многочлен Лагранжа (8.8).

Продолжая так рассуждать и далее, приходим к рекуррентному заданию последовательности интерполяционных многочленов Лагранжа, которое и составляет суть так называемой ***интерполяционной схемы Эйткена***:

$$f(x) \approx P_{0,1,\ldots,i}(x) = \frac{1}{x_i - x_0}\begin{vmatrix} x - x_0 & P_{0,1,\ldots,i-1}(x) \\ x - x_i & P_{1,2,\ldots,i}(x) \end{vmatrix}, \quad (8.17)$$

где $i = 1, 2, \ldots, n$ и, по определению, $P_0(x) := y_0$, $P_1(x) := y_1$.

Методом математической индукции можно показать идентичность (8.17) при $i = n$ и многочлена Лагранжа (8.6).

Пример 8.2. Пусть некоторая функция $y = y(x)$ задана таблицей своих значений, округленных до двух знаков после запятой:

x_i	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4
y_i	0.00	0.45	0.63	0.77	0.89	1.00	1.10	1.18

Рассмотрим поведение процесса вычисления двух значений этой функции по схеме Эйткена в точках: а) $\tilde{x} = 0.1$; б) $\tilde{x} = 0.5$. Результаты промежуточных вычислений (в которых один знак — запасной) сведем в табл.8.2 и 8.3 для случаев а) и б) соответственно. Числа в столбцах, помеченных посредством $P_{\overline{i,i+k}}(\tilde{x})$, представляют собой значения многочленов Лагранжа k-й степени, построенных по узлам от i-го до $(i+k)$-го рекуррентно по формуле

$$P_{\overline{i,i+k}}(\tilde{x}) = \frac{1}{x_{i+k} - x_i}\begin{vmatrix} \tilde{x} - x_i & P_{\overline{i,i+k-1}}(\tilde{x}) \\ \tilde{x} - x_{i+k} & P_{\overline{i+1,i+k}}(\tilde{x}) \end{vmatrix}, \quad (8.18)$$

где $k = 1, 2, \ldots$; $P_{\overline{i,i}} := y_i$, в соответствии с интерполяционной схемой Эйткена (8.17).

Таблица 8.2

Последовательность значений интерполяционных многочленов, участвующих в схеме Эйткена при вычислении значения $y(0.1)$

i	x_i	y_i	$P_{\overline{i,i+1}}(0.1)$	$P_{\overline{i,i+2}}(0.1)$	$P_{\overline{i,i+3}}(0.1)$	$P_{\overline{i,i+4}}(0.1)$	$P_{\overline{i,i+5}}(0.1)$	$P_{\overline{i,i+6}}(0.1)$	$P_{\overline{i,i+7}}(0.1)$
0	0	0.00	[1] 0.225	[3] 0.259	[6] 0.273	[10] 0.281	[15] 0.287	[21] 0.291	[28] 0.246
1	0.2	0.45	[2] 0.360	[5] 0.345	[9] 0.339	[14] 0.336	[20] 0.336	[27] 0.336	
2	0.4	0.63	[4] 0.420	[8] 0.383	[13] 0.362	[19] 0.338	[26] 0.339		
3	0.6	0.77	[7] 0.470	[12] 0.426	[18] 0.426	[25] 0.336			
4	0.8	0.89	[11] 0.505	[17] 0.426	[24] 0.570				
5	1	1.00	[16] 0.550	[23] 0.303		$y(0.1) \approx P_{\overline{0,6}}(0.1) \approx 0.29$			
6	1.2	1.10	[22] 0.660						
7	1.4	1.18							

Таблица 8.3

Вычисление по схеме Эйткена значения $y(0.5)$

i	x_i	y_i	$P_{\overline{i,i+1}}(0.5)$	$P_{\overline{i,i+2}}(0.5)$	$P_{\overline{i,i+3}}(0.5)$	$P_{\overline{i,i+4}}(0.5)$	$P_{\overline{i,i+5}}(0.5)$
0	0	0.00					
1	0.2	0.45					
2	0.4	0.63	[1] 0.700	[3] 0.700			
3	0.6	0.77	[2] 0.700				
4	0.8	0.89					
5	1	1.00					
6	1.2	1.10		$y(0.5)\approx P_{\overline{2,3}}(0.5)\approx P_{\overline{2,4}}(0.5)\approx 0.70$			
7	1.4	1.18					

Организация вычислений по формуле (8.18) должна быть такова, что если заранее неизвестна степень интерполяционного многочлена, который следует использовать для вычисления $y(\widetilde{x})$, и данная таблица значений функции достаточно обширна, то должно происходить постепенное повышение степени k интерполирующих ее многочленов за счет подключения новых, все более удаленных от $\widetilde{x}$ узлов. Порядок заполнения клеток табл.8.2 и 8.3 получаемыми по формуле (8.18) числами показан проставленными в этих клетках номерами. Счет ведется до тех пор, пока идет уточнение приближенного значения $y(\widetilde{x})$, о чем можно судить по уменьшению величины $\left|P_{\overline{i,i+k-1}}(\widetilde{x})-P_{\overline{i,i+k}}(\widetilde{x})\right|$ при увеличении k и подходящем фиксировании i.

Так, для подсчета приближенного значения данной функции в точке $\widetilde{x}=0.1$, расположенной между узлами $x_0=0$ и $x_1=0.2$, целесообразно в качестве основной последовательности значений интерполяционных многочленов Лагранжа брать числа $P_{\overline{0,1}}(0.1)=0.225=L_1(0.1)$, $P_{\overline{0,2}}(0.1)=0.259=L_2(0.1)$ и т.д., т.е. строку табл.8.2, соответствующую значению $i=0$. Составив разности между последующими и предыдущими числами этой строки, а именно:

$$0.034 \quad 0.014 \quad 0.008 \quad 0.006 \quad 0.004 \quad -0.045,$$

видим, что дальнейший счет бессмыслен; нарушается каноническое развитие процесса итерационного уточнения, и по данной информации о функции $y(x)$ более точное значение $y(0.1)$, чем значение $y(0.1)\approx 0.29$,

получаемое с помощью $L_6(0.1)$, найти не удастся*).

В случае б) для вычисления значения $y(0.5)$ оказалось достаточным сделать лишь линейную интерполяцию по двум ближайшим к точке $\widetilde{x}=0.5$ узлам $x_2=0.4$ и $x_3=0.6$ и с помощью квадратичной интерполяции по узлам $x_2=0.4$, $x_3=0.6$ и $x_4=0.8$ убедиться, что полученное линейной интерполяцией значение $y(0.1)\approx 0.70$ при этом не изменилось (см. табл.8.3).

Столь большое различие в степенях интерполяционных многочленов и, соответственно, в объеме вычислительной работы при нахождении двух промежуточных значений одной и той же функции с помощью интерполяционной схемы Эйткена нетрудно объяснить, зная, что эта функция есть $y=\sqrt{x}$. Если на левом конце промежутка $[0,1.4]$, на котором она задана своими отдельными приближенными значениями, ее производные обращаются в бесконечность, что не позволяет эффективно пользоваться формулой остаточного члена (8.12) для значения $\widetilde{x}=0.1$, близкого к $x_0=0$, то при вычислении $y=\sqrt{x}$ в точке 0.5 можно считать, что интерполяция производится на промежутке $[0.4,1.4]$, где $M_2=\max\left(\sqrt{x}\right)''\approx 1$, $M_3=\max\left(\sqrt{x}\right)'''<4$ и соответственно,

$$|R_1(0.5)|\le\frac{M_2}{2!}|\Pi_2(0.5)|\approx\frac{1}{2}|(0.5-0.4)(0.5-0.6)|=0.005,$$

$$|R_2(0.5)|\le\frac{M_3}{3!}|\Pi_3(0.5)|<\frac{4}{6}|(0.5-0.4)(0.5-0.6)(0.5-0.8)|=0.002.$$

Последняя оценка говорит о том, что квадратичная интерполяция позволила бы вычислить $\sqrt{0.5}$ с точностью практически до третьего знака после запятой, если бы данные в исходной таблице значения функции $y=\sqrt{x}$ имели такую же или бóльшую точность.

Наконец, заметим, что при компьютерных вычислениях заполнение таблиц типа табл.8.2, 8.3, разумеется, не требуется, хотя визуализация этих промежуточных данных много дает для понимания процессов интерполирования. Важно при реализации схемы Эйткена предусмотреть, например, возможность подключения узлов не только последующих, но и предшествующих заданному значению аргумента (из-за конечности таблицы исходных значений функции), причем в средней части таблицы целесообразно чередовать использование информации в последующих и предшествующих узлах.

*) На один полный итерационный шаг можно было сделать меньше, т.е. ограничиться значением $L_5(0.1)$, если учесть, что разница между $P_{\overline{0,5}}(0.1)$ и $P_{\overline{0,6}}(0.1)$ составляет величину 0.004, меньшую, чем величина ошибки округления исходных данных (0.005).

8.4. КОНЕЧНЫЕ РАЗНОСТИ

Зададимся целью придать интерполяционной формуле более простой вид, подобный виду широко используемой в математическом анализе формулы Тейлора. Если в интерполяционном многочлене Лагранжа (8.6) все слагаемые однотипны и играют одинаковую роль в образовании результата, хотелось бы иметь такое представление интерполяционного многочлена, в котором, как и в многочлене Тейлора, слагаемые располагались бы в порядке убывания их значимости. Такая структура интерполяционного многочлена позволила бы более просто перестраивать его степень, добавляя или отбрасывая удаленные от начала его записи члены.

Поставленной цели будем добиваться сначала для несколько суженной постановки задачи интерполяции. А именно, будем считать, что интерполируемая функция $y = f(x)$ задана своими значениями $y_0, y_1, \ldots, y_n$ на системе ***равноотстоящих узлов*** $x_0, x_1, \ldots, x_n$, т.е. таких, что любой узел x_i этой ***сетки*** можно представить в виде

$$x_i = x_0 + ih,$$

где $i = 0, 1, \ldots, n$, а $h > 0$ — некоторая постоянная величина, называемая ***шагом сетки*** (таблицы).

Прежде чем строить желаемые интерполяционные формулы, рассмотрим **элементы теории конечных разностей**.

Вычитая из каждого последующего члена конечной последовательности из $n+1$ чисел $y_0, y_1, \ldots, y_n$ предыдущий, образуем n ***конечных разностей первого порядка***

$$\Delta y_0 := y_1 - y_0, \quad \Delta y_1 := y_2 - y_1, \quad \ldots \quad , \quad \Delta y_{n-1} := y_n - y_{n-1}$$

или, проще, n ***первых разностей*** данной табличной функции. Из них, в свою очередь, таким же образом можно получить $n-1$ ***конечных разностей второго порядка***, или ***вторых разностей***:

$$\Delta^2 y_0 := \Delta y_1 - \Delta y_0, \quad \Delta^2 y_1 := \Delta y_2 - \Delta y_1, \quad \ldots \quad ,$$
$$\Delta^2 y_{n-2} := \Delta y_{n-1} - \Delta y_{n-2}.$$

Этот процесс построения разностей может быть продолжен, и весь он, очевидно, описывается одной рекуррентной формулой, выражающей ***конечную разность*** k***-го порядка*** $\Delta^k y_i$ через разности $(k-1)$-го порядка:

$$\Delta^k y_i = \Delta^{k-1} y_{i+1} - \Delta^{k-1} y_i, \tag{8.19}$$

где $k = 1, 2, \ldots, n$ и $\Delta^0 y_i := y_i$.

В некоторых случаях требуется знать **выражения конечных разностей непосредственно через значения функции**, лежащей в их основе. Для нескольких первых порядков разностей их можно получить прямой подстановкой:

$$\Delta y_i = y_{i+1} - y_i,$$

$$\Delta^2 y_i = \Delta y_{i+1} - \Delta y_i = y_{i+2} - y_{i+1} - (y_{i+1} - y_i) = y_{i+2} - 2y_{i+1} + y_i,$$

$$\Delta^3 y_i = \Delta^2 y_{i+1} - \Delta^2 y_i = y_{i+3} - 2y_{i+2} + y_{i+1} - (y_{i+2} - 2y_{i+1} + y_i) =$$
$$= y_{i+3} - 3y_{i+2} + 3y_{i+1} - y_i, \quad \text{и т.д.}$$

Подметив закономерность в коэффициентах рассмотренных представлений конечных разностей, записываем общую формулу

$$\Delta^k y_i = \sum_{j=0}^{k} (-1)^j C_k^j y_{k+i-j}, \tag{8.20}$$

которая может быть строго обоснована методом математической индукции и которая напоминает биномиальное разложение для $(y-1)^k$.

Привлекая определение производной, можно обнаружить прямую **связь между конечными разностями и производными**. А именно, если учесть, что

$$\lim_{h\to 0} \frac{y_{i+1} - y_i}{h} = \lim_{h\to 0} \frac{f(x_i + h) - f(x_i)}{h} = f'(x_i),$$

то можно сказать, что при малых h имеет место приближенное равенство

$$\Delta y_i \approx f'(x_i)h,$$

т.е. первые разности характеризуют первую производную функции $f(x)$, по значениям которой они составлены. Пользуясь этим, имеем для вторых разностей:

$$\frac{\Delta^2 y_i}{h^2} = \frac{\Delta y_{i+1} - \Delta y_i}{h^2} = \frac{\dfrac{y_{i+2} - y_{i+1}}{h} - \dfrac{y_{i+1} - y_i}{h}}{h} \approx$$
$$\approx \frac{f'(x_i + h) - f'(x_i)}{h} \approx f''(x_i),$$

т.е. $\Delta^2 y_i \approx f''(x_i)h^2$, и, вообще,

$$\Delta^k y_i \approx f^{(k)}(x_i)h^k. \tag{8.21}$$

Таким образом, на конечные разности можно смотреть как на некоторый аналог производных*). Отсюда справедливость многих их свойств, одинаковых со свойствами производных.

*) Более подробно связь между производными и конечными разностями изучается далее в главе 13.

Отметим лишь **простейшие свойства конечных разностей**:

1) *конечные разности постоянной равны нулю* (очевидно);

2) *постоянный множитель у функции можно выносить за знак конечной разности.*

Действительно,

$$\Delta(Cy(x)) = Cy(x+h) - Cy(x) = C[y(x+h) - y(x)] = C\Delta y(x)$$

при любых фиксированных x и постоянной C;

3) *конечная разность от суммы двух функций равна сумме их конечных разностей в одной и той же точке.*

Свойство проверяется непосредственно: при любых x

$$\Delta(u(x)+v(x)) = u(x+h) + v(x+h) - (u(x)+v(x)) =$$
$$= u(x+h) - u(x) + v(x+h) - v(x) = \Delta u(x) + \Delta v(x).$$

Свойства 2 и 3 характеризуют операцию взятия конечной разности как линейную операцию.

Учитывая роль, которую играют многочлены в теории интерполирования, посмотрим, что представляют собой **конечные разности многочлена**.

Поскольку многочлен в своей канонической форме есть линейная комбинация степенных функций, положим сначала $y = x^n$. Используя биномиальное разложение n-й степени двучлена, получим:

$$\Delta\left(x^n\right) = (x+h)^n - x^n = nhx^{n-1} + \frac{n(n-1)}{2!}h^2x^{n-2} + \ldots + nh^{n-1}x + h^n,$$

т.е. первая конечная разность степенной функции $y = x^n$ есть многочлен степени $n-1$ со старшим членом nhx^{n-1}. Если взять теперь конечную разность от функции

$$y = a_0x^n + a_1x^{n-1} + \ldots + a_{n-1}x + a_n, \qquad (8.22)$$

то, в силу линейных свойств Δy, можно записать

$$\Delta y(x) = a_0\,\Delta\left(x^n\right) + a_1\,\Delta\left(x^{n-1}\right) + \ldots + a_{n-1}\,\Delta x.$$

Первое слагаемое в этой сумме, как выяснено, есть многочлен $(n-1)$-й степени, второе, аналогично, — многочлен степени $n-2$, и т.д. Следовательно, первая конечная разность многочлена (8.22) в точке x с шагом h есть тоже многочлен со старшим членом a_0nhx^{n-1}, вторая конечная разность — многочлен со старшим членом $a_0n(n-1)h^2x^{n-2}$, ..., k-я разность — многочлен со старшим членом $a_0n(n-1)\ldots(n-k+1)h^kx^{n-k}$.

При $k = n$ получаем п о с т о я н н у ю разность n-го порядка

$$\Delta^n y = a_0 n! h^n$$

для многочлена (8.22); конечные разности более высоких порядков, естественно, равны нулю.

Итак, главный вывод из предыдущих рассуждений: *n-е конечные разности многочлена n-й степени постоянны, а $(n+1)$-е и все последующие равны нулю.*

Более важным для понимания сути полиномиального интерполирования является утверждение, обратное сделанному выше выводу. А именно, доказано [58, 123], что *если конечные разности n-го порядка некоторой функции $y = y(x)$ постоянны в любой точке x при различных фиксированных шагах h, то эта функция $y(x)$ есть многочлен степени n.*

Для функции $y = f(x)$, заданной таблицей своих значений $y_0, y_1, \ldots, y_n$ в узлах $x_0, x_1, \ldots, x_n$, где $x_i = x_0 + ih$, конечные разности разных порядков удобно помещать в одну общую таблицу с узлами и значениями функции (последние можно интерпретировать как конечные разности нулевого порядка, см. (8.19)). Эту общую таблицу называют ***таблицей конечных разностей***. Заметим, что кроме принятого здесь так называемого диагонального расположения конечных разностей, когда числа в каждом столбце записываются со смещением на полстроки так, как это показано в табл.8.4, часто применяют горизонтальное расположение, где Δy_i, $\Delta^2 y_i$ и другие разности с индексом i помещают в одной строке с x_i, y_i.

Т а б л и ц а 8.4

Диагональная таблица конечных разностей

x_0	y_0					
		Δy_0				
x_1	y_1		$\Delta^2 y_0$			
		Δy_1		$\Delta^3 y_0$		
x_2	y_2		$\Delta^2 y_1$		$\Delta^4 y_0$	
		Δy_2		$\Delta^3 y_1$		...
x_3	y_3		$\Delta^2 y_2$		$\Delta^4 y_1$	
		Δy_3		$\Delta^3 y_2$		
x_4	y_4		$\Delta^2 y_3$		$\vdots$	
		Δy_4		$\vdots$		
x_5	y_5		$\vdots$			
$\vdots$	$\vdots$	$\vdots$				

Таблица 8.5

Конечные разности функции $y = x\ln^2 x$

x_i	y_i	Δy_i	$\Delta^2 y_i$	$\Delta^3 y_i$	$\Delta^4 y_i$	$\Delta^5 y_i$	$\Delta^6 y_i$	$\Delta^7 y_i$	$\Delta^8 y_i$	$\Delta^9 y_i$	$\Delta^{10} y_i$
0.4	0.336	−0.179	0.062	0.015	−0.012	0.008	−0.007	0.013	−0.035	0.090	−0.207
0.6	0.157	−0.117	0.077	0.003	−0.004	0.001	0.006	−0.022	0.055	−0.117	
0.8	0.040	−0.040	0.080	−0.001	−0.003	0.007	−0.016	0.033	−0.062		
1.0	0.000	0.040	0.079	−0.004	0.004	−0.009	0.017	−0.029			
1.2	0.040	0.119	0.075	0.000	−0.005	0.008	−0.012				
1.4	0.159	0.194	0.075	−0.005	0.003	−0.004					
1.6	0.353	0.269	0.070	−0.002	−0.001						
1.8	0.622	0.339	0.068	−0.003							
2.0	0.961	0.407	0.065								
2.2	1.368	0.472									
2.4	1.840										

Пример 8.3. Составим таблицу конечных разностей для функции $y = x\ln^2 x$ по ее значениям, вычисленным с тремя знаками после запятой в точках $x_i = 0.4 + 0.2i$, где $i = 0, 1, \ldots, 10$. В соответствии с формой, задаваемой табл.8.4, заполняем табл.8.5 всеми возможными для этого случая конечными разностями. Проанализируем ее.

Учитывая связь между конечными разностями и производными соответствующих порядков (см. (8.21)), по смене знаков чисел в столбце Δy_i можно судить о наличии минимума функции в окрестности точки $x = 1$, а положительность всех чисел в столбце $\Delta^2 y_i$ говорит о выпуклости вниз графика данной функции на всем рассматриваемом промежутке $[0.4, 2.4]$.

Далее замечаем, что абсолютные величины конечных разностей сначала убывают с увеличением их порядка, а затем начинают увеличиваться. Это типичное поведение конечных разностей при ограниченной точности задания значений сеточной функции.

Природу наблюдаемого в примере 8.3 поведения модулей конечных разностей нетрудно понять. Если шаг достаточно мал, а данная табличная функция — достаточно гладкая, то сначала происходит естественное убывание $|\Delta^k y_i|$ с увеличением k, в силу упомянутой связи (8.21). Когда эти величины становятся достаточно малыми, бóльшую роль начинают играть продукты взаимодействия исходных ошибок округления (так называемый ***шум округлений***).

Что происходит с одной отдельно взятой ошибкой величины ε у значения y_i, можно проследить по табл.8.6. Как видим, с ростом порядка разностей она «расползается» по таблице и увеличивается по абсолютной величине.

Таблица 8.6

Продвижение ошибки по таблице конечных разностей

…	…	…	…	…	…	
x_{i-2}	y_{i-2}		$\Delta^2 y_{i-3}$		$\Delta^4 y_{i-4} + \varepsilon$	
		Δy_{i-2}		$\Delta^3 y_{i-3} + \varepsilon$		
x_{i-1}	y_{i-1}		$\Delta^2 y_{i-2} + \varepsilon$		$\Delta^4 y_{i-3} - 4\varepsilon$	
		$\Delta y_{i-1} + \varepsilon$		$\Delta^3 y_{i-2} - 3\varepsilon$		…
x_i	$y_i + \varepsilon$		$\Delta^2 y_{i-1} - 2\varepsilon$		$\Delta^4 y_{i-2} + 6\varepsilon$	
		$\Delta y_i - \varepsilon$		$\Delta^3 y_{i-1} + 3\varepsilon$		…
x_{i+1}	y_{i+1}		$\Delta^2 y_i + \varepsilon$		$\Delta^4 y_{i-1} - 4\varepsilon$	
		Δy_{i+1}		$\Delta^3 y_i - \varepsilon$		
x_{i+2}	y_{i+2}		$\Delta^2 y_{i+1}$		$\Delta^4 y_i + \varepsilon$	
…	…	…	…	…	…	

Погрешности, имеющиеся у каждого из данных значений функции, с ростом порядка разностей все больше взаимодействуют.

Из сделанных наблюдений напрашивается следующий вывод. Если какой-то столбец в таблице конечных разностей (в ее эксплуатируемой части) состоит из чисел, абсолютные величины которых составляют всего несколько единиц десятичного знака, являющегося последним в записи исходных значений функции, скажем, не превосходят величины 10ε, где ε — абсолютная погрешность исходных данных, то эти конечные разности и разности всех последующих порядков не несут практически никакой информации о функции, и их не следует использовать. Разности же предшествующего столбца называются ***практически постоянными***, и их порядок определяет степень многочлена, которую можно и должно использовать для идеальной в данных условиях полиномиальной интерполяции.

Вспоминая о том, что многочлен k-й степени имеет k-е разности постоянными, а все последующие — нулевыми, приходим к заключению, что *если k-е разности таблицы конечных разностей некоторой функции практически постоянны, то эта функция ведет себя в рассматриваемой области, как многочлен k-й степени*; эту степень и следует применять для интерполирования с наибольшей для данных реалий точностью.

Обратимся к числовой табл.8.5 нашего примера. Видим, что если исключить из рассмотрения верхнюю диагональную строку, то для всей остальной части таблицы третьи разности удовлетворят условию $|\Delta^3 y_i| \le 10 \cdot 0.0005$ (где 0.0005 — предельная абсолютная погрешность значений y_i). В такой ситуации разности более высоких порядков не следовало вообще вычислять, а разности второго порядка можно считать практически постоянными, т.е. для подсчета любых промежуточных значений данной функции, за исключением, быть может, тех, которые находятся вблизи узла $x_0 = 0.4$, нужно применять квадратичную интерполяцию.

8.5. КОНЕЧНОРАЗНОСТНЫЕ ИНТЕРПОЛЯЦИОННЫЕ ФОРМУЛЫ

Пусть функция $y = f(x)$ задана на сетке равноотстоящих узлов $x_i = x_0 + ih$, где $i = 0, 1, \ldots, n$, и для нее построена таблица конечных разностей 8.4.

В соответствии с тем, что было сказано о направлении модификации интерполяционной формулы Лагранжа в начале предыдущего параграфа, будем строить интерполяционный многочлен $P_n(x)$ в форме

$$P_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \ldots \\ \ldots + a_n(x - x_0)(x - x_1)\ldots(x - x_{n-1}). \quad (8.23)$$

Его $n+1$ коэффициент $a_0, a_1, \ldots, a_n$ будем находить последовательно из $n+1$ интерполяционных равенств

$$P_n(x_i) = y_i, \qquad i = 0, 1, \ldots, n.$$

А именно, полагая $i = 0$, т.е. $x = x_0$, в (8.23) имеем $P_n(x_0) = a_0$, а по условию интерполяции $P_n(x_0) = y_0$; следовательно, $a_0 = y_0$.

Далее, при $i = 1$ аналогично получаем равенство

$$a_0 + a_1(x_1 - x_0) = y_1,$$

в которое подставляем уже найденное значение $a_0 = y_0$. Разрешая это равенство относительно a_1 и используя обозначение конечной разности, получаем

$$a_1 = \frac{y_1 - y_0}{x_1 - x_0} = \frac{\Delta y_0}{h}.$$

Следующий шаг, при $i = 2$, дает:

$$a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1) = y_2 \quad \Leftrightarrow$$

$$y_0 + \frac{\Delta y_0}{h} \cdot 2h + a_2 \cdot 2h \cdot h = y_2 \quad \Leftrightarrow$$

$$a_2 = \frac{y_2 - 2y_1 + y_0}{2!h^2} = \frac{\Delta^2 y_0}{2!h^2} \qquad \text{(см. (8.20) при } k = 2).$$

Полной индукцией можно показать справедливость выражения

$$a_k = \frac{\Delta^k y_0}{k!h^k} \qquad \forall\, k \in \{1, 2, \ldots, n\}. \quad (8.24)$$

Подставляя найденные коэффициенты $a_0, a_1, \ldots, a_n$ в (8.23), получаем многочлен

$$P_n(x) = y_0 + \frac{\Delta y_0}{h}(x - x_0) + \frac{\Delta^2 y_0}{2!h^2}(x - x_0)(x - x_1) + \ldots \\ \ldots + \frac{\Delta^n y_0}{n!h^n}(x - x_0)(x - x_1)\ldots(x - x_{n-1}), \quad (8.25)$$

который называют ***первым интерполяционным многочленом Ньютона***.

12 Б-223

Учитывая, что каждое слагаемое многочлена (8.25), начиная со второго, содержит множитель $x - x_0$, естественно предположить, что этот многочлен наиболее приспособлен для интерполирования в окрестности узла x_0 (при x, близких к x_0, $f(x) \approx y_0$). Будем называть узел x_0 ***базовым*** для многочлена (8.25), и упростим (8.25) введением новой переменной q равенством $q = \dfrac{x - x_0}{h}$, или (что то же) равенством $x = x_0 + qh$. Так как при любых $i \in \{0, 1, \ldots, n\}$

$$x - x_i = x_0 + qh - x_0 - ih = h(q - i),$$

то в результате подстановки этих разностей в (8.25) приходим к ***первой интерполяционной формуле Ньютона*** в виде

$$f(x) \approx P_n(x_0 + qh) := y_0 + q\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_0 + \ldots$$
$$\ldots + \frac{q(q-1)\ldots(q-n+1)}{n!}\Delta^n y_0, \quad (8.26)$$

где обозначение $P_n(x_0 + qh)$ указывает не только на n-ю степень многочлена, но и на базовый узел x_0 и связь переменных x и q.

Первая формула Ньютона (8.26) обычно применяется при значениях $|q| < 1$, а именно, *для интерполирования вперед*, (при $x \in (x_0, x_1)$, т.е. при $q \in (0, 1)$) *и экстраполирования назад* (при $x < x_0$, т.е. при $q < 0$).

Так как реально степени интерполяционных многочленов бывают не так велики, в то время как таблицы значений функций достаточно обширны, и так как в реальной числовой таблице никаких индексов — номеров узлов нет (см., например, приведенную выше табл.8.5), то за базовый для формулы (8.26) узел x_0 можно принимать узел, ближайший к заданной фиксированной точке x, если за ним имеется достаточное число узлов для построения необходимых для (8.26) разностей. Поскольку в первой формуле Ньютона используются нисходящие диагонали таблицы конечных разностей (см. табл.8.4), то такое смещение узла, принимаемого за базовый, в конце таблицы будет неприемлемо.

Учет этого обстоятельства приводит к потребности в симметричной, в определенном смысле, для (8.26) формулы, которая была бы пригодной для интерполирования в конце таблицы. Для этого, в отличие от (8.23), форма интерполяционного многочлена $P_n(x)$ берется такой, которая предусматривает поочередное подключение узлов в обратном порядке: сначала последний, потом

предпоследний и т.д., т.е.

$$P_n(x) = a_0 + a_1(x - x_n) + a_2(x - x_n)(x - x_{n-1}) + \ldots$$
$$\ldots + a_n(x - x_n)(x - x_{n-1})\ldots(x - x_1).$$

Коэффициенты $a_0, a_1, \ldots, a_n$ этого многочлена находятся аналогично тому, как они находились для многочлена (8.23), только здесь подстановка узловых точек вместо x и рассмотрение интерполяционных равенств производится тоже в обратном порядке. Полагая $x = x_n$, $x = x_{n-1}, \ldots$, имеем:

$$P_n(x_n) = a_0 = y_n,$$

$$P_n(x_{n-1}) = y_n + a_1(x_{n-1} - x_n) = y_{n-1} \Rightarrow a_1 = \frac{y_n - y_{n-1}}{x_n - x_{n-1}} = \frac{\Delta y_{n-1}}{h},$$

$$P_n(x_{n-2}) = y_n + \frac{\Delta y_{n-1}}{h}(x_{n-2} - x_n) + a_2(x_{n-2} - x_n)(x_{n-2} - x_{n-1}) = y_{n-2}$$

$$\Rightarrow \quad a_2 = \frac{y_{n-2} - y_n + 2\Delta y_{n-1}}{(x_{n-2} - x_n)(x_{n-2} - x_{n-1})} = \frac{y_n - 2y_{n-1} + y_{n-2}}{-2h(-h)} = \frac{\Delta^2 y_{n-2}}{2!h^2}$$

и т.д. В общем случае

$$a_k = \frac{\Delta^k y_{n-k}}{k!h^k} \qquad \forall\, k \in \{1, 2, \ldots, n\}.$$

Таким образом получаем ***второй интерполяционный многочлен Ньютона***

$$P_n(x) = y_n + \frac{\Delta y_{n-1}}{h}(x - x_n) + \frac{\Delta^2 y_{n-2}}{2!h^2}(x - x_n)(x - x_{n-1}) + \ldots$$
$$\ldots + \frac{\Delta^n y_0}{n!h^n}(x - x_n)(x - x_{n-1})\ldots(x - x_1), \qquad (8.27)$$

в котором базовым является узел x_n и коэффициенты которого определяются конечными разностями, расположенными на восходящей от y_n диагонали.

Положим в (8.27) $x = x_n + qh$, иначе, введем новую переменную $q = \dfrac{x - x_n}{h}$ и преобразуем к ней входящие в (8.27) разности:

$$x - x_i = x_n + qh - x_0 - ih = x_0 + nh + qh - x_0 - ih = h(q + n - i).$$

В результате приходим ко ***второй интерполяционной формуле Ньютона*** вида

$$f(x) \approx P_n(x_n + qh) = y_n + q\Delta y_{n-1} + \frac{q(q+1)}{2!}\Delta^2 y_{n-2} + \ldots$$
$$\ldots + \frac{q(q+1)\ldots(q+n-1)}{n!}\Delta^n y_0. \quad (8.28)$$

Ее также целесообразно использовать при значениях $|q| < 1$, т.е. в окрестности узла x_n *для интерполирования назад* (при $q \in (-1, 0)$) *и экстраполирования вперед* (при $q > 0$).

Наряду с выведенными специально для начала и конца таблицы первой и второй интерполяционными формулами Ньютона, имеется еще несколько формул, рассчитанных на их применение в центральной части таблицы и потому называемых ***центральными интерполяционными формулами***. Прежде, чем определять эти формулы, введем понятие центральных разностей.

Будем считать, что узел x_0 расположен в середине таблицы, и нумерация остальных узлов производится, начинаясь с x_0, с использованием как положительных, так и отрицательных индексов, т.е. считаем $x_i = x_0 + ih$, где $i = 0, \pm 1, \pm 2, \ldots$. Тогда центральная часть таблицы конечных разностей будет проиндексирована так, как это показано в табл.8.7. Все подчеркнутые в ней конечные разности (находящиеся с x_0, y_0 в одной строке и на полстроки выше и ниже) называются ***центральными разностями***.

Интерполяционный многочлен ищем в форме

$$P(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) +$$
$$+ a_3(x - x_{-1})(x - x_0)(x - x_1) +$$
$$+ a_4(x - x_{-1})(x - x_0)(x - x_1)(x - x_2) + \ldots, \quad (8.29)$$

предполагающей постепенное подключение узлов x_i: сначала при $i = 0$, затем при $i = 1$, потом при $i = -1$ и т.д., т.е. с двух сторон от x_0. При этом здесь и далее не будем фиксировать степени многочленов и не будем стремиться выписывать общие и, тем более, последние члены таких многочленов. Как и в предыдущих случаях, коэффициенты a_k $(k = 0, 1, 2, \ldots)$ находим один за другим последовательной подстановкой в $P(x)$ и в интерполяционные равенства $P(x_i) = y_i$ значений $x = x_0, x_1, x_{-1}, x_2, x_{-2}, \ldots$:

$$a_0 = y_0; \quad a_1 = \frac{\Delta y_0}{h}; \quad a_2 = \frac{\Delta^2 y_{-1}}{2!h^2}; \quad a_3 = \frac{\Delta^3 y_{-1}}{3!h^3}; \quad a_4 = \frac{\Delta^4 y_{-2}}{4!h^4}$$

и т.д. Введя новую переменную $q = \dfrac{x - x_0}{h}$ и выразив через нее

разности $x - x_i = h(q - i)$ для всех $i = 0, \pm 1, \pm 2, \ldots$, в результате подстановки этих разностей и выражений коэффициентов в шаблон (8.29), приходим к ***первой интерполяционной формуле Гаусса***:

$$f(x) \approx \overline{P}(x_0 + qh) = y_0 + q\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_{-1} +$$
$$+ \frac{(q+1)q(q-1)}{3!}\Delta^3 y_{-1} + \frac{(q+1)q(q-1)(q-2)}{4!}\Delta^4 y_{-2} + \ldots. \quad (8.30)$$

Записанные слагаемые легко дополнить следующими, если знать, что в этой формуле используются нижние центральные разности все возрастающих порядков, т.е. те, которые подчеркнуты в табл.8.7 сплошной чертой.

Таблица 8.7

Таблица центральных разностей

x_i	y_i	Δy_i	$\Delta^2 y_i$	$\Delta^3 y_i$	$\Delta^4 y_i$	$\Delta^5 y_i$	$\Delta^6 y_i$	...
...	...							
		...						
x_{-3}	y_{-3}		...					
		Δy_{-3}		...				
x_{-2}	y_{-2}		$\Delta^2 y_{-3}$		...			
		Δy_{-2}		$\Delta^3 y_{-3}$		...		
x_{-1}	y_{-1}		$\Delta^2 y_{-2}$		$\Delta^4 y_{-3}$		...	
		Δy_{-1}		$\Delta^3 y_{-2}$		$\Delta^5 y_{-3}$		
x_0	y_0		$\underline{\Delta^2 y_{-1}}$		$\underline{\Delta^4 y_{-2}}$		$\underline{\Delta^6 y_{-3}}$	...
		$\underline{\Delta y_0}$		$\underline{\Delta^3 y_{-1}}$		$\underline{\Delta^5 y_{-2}}$		
x_1	y_1		$\Delta^2 y_0$		$\Delta^4 y_{-1}$		...	
		Δy_1		$\Delta^3 y_0$		...		
x_2	y_2		$\Delta^2 y_1$		...			
		Δy_2		...				
x_3	y_3		...					
		...						
...	...							

Совершенно аналогично, подключая узлы в другом порядке (после x_0 сначала предшествующий, затем последующий и т.д.,

т.е. $x_0, x_{-1}, x_1, x_{-2}, \ldots$), можно построить ***вторую интерполяционную формулу Гаусса***

$$f(x) \approx \widetilde{P}(x_0 + qh) = y_0 + q\Delta y_{-1} + \frac{(q+1)q}{2!}\Delta^2 y_{-1} +$$

$$+\frac{(q+1)q(q-1)}{3!}\Delta^3 y_{-2} + \frac{(q+2)(q+1)q(q-1)}{4!}\Delta^4 y_{-2} + \ldots, \qquad (8.31)$$

использующую в е р х н и е центральные разности (подчеркнутые в табл.8.7 пунктирной линией).

Интерполяционные формулы Гаусса служат полуфабрикатами для получения более симметричных, использующих в с е центральные разности интерполяционных формул.

Так, полусумма первого и второго интерполяционных многочленов Гаусса после преобразований приводит к формуле

$$f(x) \approx P_S(x_0 + qh) = y_0 + q\frac{\Delta y_{-1} + \Delta y_0}{2} + \frac{q^2}{2!}\Delta^2 y_{-1} +$$

$$+\frac{q(q^2-1)}{3!}\cdot\frac{\Delta^3 y_{-2} + \Delta^3 y_{-1}}{2} + \frac{q^2(q^2-1)}{4!}\Delta^4 y_{-2} + \ldots, \qquad (8.32)$$

называемой ***интерполяционной формулой Стирлинга****).

Если же взять полусумму второго интерполяционного многочлена Гаусса и такого же многочлена, но с нижними индексами, увеличенными на единицу (т.е. с базовой точкой x_1 вместо x_0), то придем к ***интерполяционной формуле Бесселя*****)

$$f(x) \approx P_B(x_0 + qh) = \frac{y_0 + y_1}{2} + \left(q - \frac{1}{2}\right)\Delta y_0 +$$

$$+\frac{q(q-1)}{2!}\cdot\frac{\Delta^2 y_{-1} + \Delta^2 y_0}{2} + \frac{\left(q-\frac{1}{2}\right)q(q-1)}{3!}\Delta^3 y_{-1} +$$

$$+\frac{q(q-1)(q+1)(q-2)}{4!}\cdot\frac{\Delta^4 y_{-2} + \Delta^4 y_{-1}}{2} + \ldots. \qquad (8.33)$$

В последней формуле обращает на себя внимание тот факт, что она сильно упростится, если в нее подставить значение $q = \frac{1}{2}$, соответствующее значению аргумента $\hat{x} = \frac{1}{2}(x_0 + x_1)$.

---

*) Стирлинг Джеймс (1692–1770) — шотландский математик.

**) Бéссель Фридрих Вильгельм (1784–1846) — немецкий астроном, геодезист и математик.

Этот частный случай формулы Бесселя называют ***формулой интерполирования на середину***:

$$f\left(\frac{x_0 + x_1}{2}\right) \approx \frac{y_0 + y_1}{2} - \frac{1}{8}\cdot\frac{\Delta^2 y_{-1} + \Delta^2 y_0}{2} +$$

$$+\frac{3}{128}\cdot\frac{\Delta^4 y_{-2} + \Delta^4 y_{-1}}{2} - \ldots. \qquad (8.34)$$

Итак, *если точка x, в которой нужно найти приближенное значение таблично заданной функции $f(x)$, находится в начале или в конце таблицы, применяется соответственно первая (8.26) или вторая (8.28) формулы Ньютона с таким выбором базовой точки, чтобы значение $|q|$ было как можно меньше. Если точка x находится в середине таблицы, то всегда можно зафиксировать точку x_0 в таблице центральных разностей так, чтобы $q = \frac{x - x_0}{h}$ либо было по модулю меньше 0.25 и тогда применять интерполяционную формулу Стирлинга (8.32), либо чтобы $q \in [0.25,\ 0.75]$ и использовать формулу Бесселя (8.33).*

Пример 8.4. Пусть требуется для функции $y = f(x)$, заданной в примере 8.3 таблицей нескольких своих значений с тремя знаками после запятой, найти приближенные значения: а) $f(0.5)$; б) $f(1.22)$; в) $f(1.5)$; г) $f(1.94)$; д) $f(2.5)$, записав предварительно соответствующие каждому случаю интерполяционные формулы.

Для решения поставленной задачи учитываем, что значения $f(x)$ заданы в примере 8.3 на сетке р а в н о о т с т о я щ и х узлов, поэтому здесь можно применить конечноразностную интерполяцию. При этом будем пользоваться уже составленной табл.8.5 конечных разностей и проведенным ранее ее анализом на выявление оптимальной степени многочлена. Для случаев б–д фиксируем вторую степень, для а — третью. В каждом случае, т.е. для конкретного значения аргумента, выбираем базовый узел, подсчитываем значение вспомогательной переменной q и, в зависимости от положения базового узла и значения q, пользуясь представленными в табл.8.5 числами, записываем требуемую интерполяционную формулу. Подстановка в нее значения q приводит к искомому значению $f(x)$.

а) При $x = 0.5$ (начало таблицы) полагаем $x_0 = 0.4$; тогда $q = \frac{x - x_0}{h} = \frac{0.5 - 0.4}{0.2} = 0.5$. Соответствующую интерполяционную формулу для аппроксимации $f(x)$ при $x = 0.4 + 0.2q$ с $q \in (-1, 1)$ записываем, глядя на первую интерполяционную формулу Ньютона (8.26) и табл.8.5:

$$f(x) \approx P_3(0.4 + 0.2q) = 0.336 - 0.179q + \frac{0.062}{2}q(q-1) + \frac{0.015}{6}q(q-1)(q-2).$$

Отсюда получаем искомое значение

$$f(0.5) \approx P_3(0.4+0.2\cdot 0.5) = 0.336-0.179\cdot 0.5+0.031\cdot 0.5\cdot(-0.5)+$$
$$+0.0025\cdot 0.5\cdot(-0.5)\cdot(-1.5) \approx 0.336-0.0895-0.0078+0.0009 \approx 0.240$$

б) Точка $x = 1.22$ находится в средней части таблицы. Поэтому здесь целесообразно применить формулу Стирлинга или Бесселя. Полагая $x_0 = 1.2$ и найдя $q = \frac{x-x_0}{h} = \frac{1.22-1.2}{0.2} = 0.1$, останавливаемся на формуле Стирлинга (8.32), которая в данном случае имеет вид

$$f(x) \approx P_S(1.2+0.2q) = 0.040 + \frac{0.040+0.119}{2}q + \frac{0.079}{2}q^2$$

и при $q = 0.1$ приводит к искомому значению

$$f(1.22) \approx 0.040 + 0.0080 + 0.0004 \approx 0.048.$$

в) Здесь, очевидно, напрашивается применение формулы (8.34) интерполирования на середину. Полагая $x_0 = 1.4$, $x_1 = 1.6$, имеем:

$$f(1.5) \approx \frac{0.159+0.353}{2} - \frac{1}{8}\cdot\frac{0.075+0.075}{2} \approx 0.256 - 0.0094 \approx 0.247.$$

г) Глядя на положение точки $x = 1.94$ в заданной системе узлов табл.8.5, видим, что для вычисления $f(1.94)$ также возможно применение центральных интерполяционных формул. Положив $x_0 = 1.8$ и вычислив $q = \frac{x-x_0}{h} = \frac{1.94-1.8}{0.2} = 0.7$, на основе (8.33) записываем интерполяционную формулу Бесселя

$$f(x) \approx P_B(1.8+0.2q) = \frac{0.622+0.961}{2} + 0.339(q-0.5) + \frac{0.070+0.068}{2}\cdot\frac{q(q-1)}{2}.$$

Из нее получаем

$$f(1.94) \approx P_B(1.8+0.2\cdot 0.7) \approx 0.7915 + 0.0678 - 0.0072 \approx 0.852.$$

д) Точка $x = 2.5$ расположена за последним узлом, поэтому для экстраполяции $f(x)$ здесь однозначно следует применить вторую интерполяционную формулу Ньютона (8.28). Считая $x_n = 2.4$ (индекс n здесь используется условно, без придания ему конкретного значения), записываем формулу экстраполяции

$$f(x) \approx P(2.4+0.2q) = 1.840 + 0.472q + \frac{0.065}{2}q(q+1),$$

откуда при $q = \frac{x-x_n}{h} = \frac{2.5-2.4}{0.2} = 0.5$ находим

$$f(2.5) \approx P(2.4+0.2\cdot 0.5) \approx 1.840 + 0.236 + 0.0244 \approx 2.100.$$

В порядке обсуждения приведенного примера отметим следующее. Во-первых, при записи интерполяционных многочленов не стоит приводить их к канонической форме, ибо тогда их члены утратят ту информативность, которая в них заложена по построению и которая хорошо видна из рассмотрения промежуточных результатов: налицо убывание роли слагаемых при подсчете каждого из значений а–д. Во-вторых, вычисление слагаемых интерполяционного результата нет смысла проводить более чем с одним запасным знаком*), который в конце должен быть отброшен. Об этом говорят приближенность исходных значений и правила приближенных вычислений [3, 98, и др.]. Точность результата интерполирования принципиально не может быть выше, чем точность исходных данных. Вообще, о точности конечноразностного интерполирования речь пойдет чуть ниже; здесь же заметим, что гладкость данной функции, плотность системы узлов и выбор подходящих параметров интерполяционных формул обеспечили в примере 8.4 вычисление всех требуемых значений, как нетрудно убедиться, с точностью не хуже 0.001, т.е. в грубом смысле все десятичные знаки результатов — верные.

Теперь о том, как могут быть трансформированы **остаточный член и его оценки при конечноразностной интерполяции**.

В силу доказанной в § 8.2 единственности интерполяционного многочлена Лагранжа, все построенные здесь конечноразностные интерполяционные многочлены Ньютона и Гаусса — это всего лишь различные формы его представления**). Следовательно, для всех этих форм справедливо выражение остаточного члена (8.12), где определенный посредством (8.10) многочлен $\Pi_{n+1}(x)$ для случая равноотстоящих узлов $\{x_i\}_{i=0}^n$ преобразуется к новой переменной $q = \frac{x-x_0}{h}$ следующим образом:

$$\Pi_{n+1}(x) = \Pi_{n+1}(x_0+qh) = \prod_{i=0}^{n}(x_0+qh-x_0-ih) = h^{n+1}\prod_{i=0}^{n}(q-i).$$

Отсюда

$$R_n(x) = R_n(x_0+qh) = \frac{f^{(n+1)}(\xi)}{(n+1)!}h^{n+1}q(q-1)\ldots(q-n). \qquad (8.35)$$

Итак, для $f(x) \in C^{n+1}[a, b]$ конечноразностная интерполяционная формула Ньютона или Гаусса степени n с базовым уз-

*) При машинном счете эта рекомендация, разумеется, утрачивает смысл.

**) По поводу формул Стирлинга и Бесселя см. далее замечание 8.1.

лом x_0 может быть записана в виде

$$f(x) = P_n(x_0 + qh) + R_n(x_0 + qh), \qquad (8.36)$$

где $P_n(x_0 + qh)$ — тот или иной конечноразностный многочлен, построенный по равноотстоящим (с шагом h) узлам $x_0, x_1, \ldots, x_n \in [a, b]$, а $R_n(x_0 + qh)$ — остаточный член (8.35), в котором ξ — некоторая неизвестная, но фиксированная (при фиксированном x) точка интервала (a, b).

Аналогично, при выборе базового узла x_n, т.е. для второй интерполяционной формулы Ньютона, получаем точное представление

$$f(x) = P_n(x_n + qh) + R_n(x_n + qh),$$

где $R_n(x_n + qh) = \dfrac{f^{(n+1)}(\xi)}{(n+1)!} q(q+1)\ldots(q+n)$, а $P_n(x_n + qh)$ — многочлен, определенный в (8.28).

При наличии оценки $\left|f^{(n+1)}(x)\right| \le M_{n+1} \quad \forall\, x \in [a, b]$ можно уточнить границы абсолютной погрешности конечноразностного интерполирования в конкретной точке и на всем промежутке $[a, b]$ по типу (8.14), (8.15).

Например, для оценки погрешности интерполяции в точке $\tilde{x} = x_0 + \tilde{q}h$ на основании (8.36), (8.35) имеем*)

$$|f(\tilde{x}) - P_n(x_0 + \tilde{q}h)| \le \frac{M_{n+1}}{(n+1)!} h^{n+1} |\tilde{q}(\tilde{q} - 1)\ldots(\tilde{q} - n)|. \qquad (8.37)$$

Если интерполяционная формула $f(x) \approx P_n(x_0 + qh)$ используется для аппроксимации $f(x)$ в точке $\tilde{x}$, расположенной достаточно близко к базовому узлу x_0 справа от него, т.е. если $\tilde{q} \in (0, 1)$, то оценку (8.37) можно существенно упростить.

*) В силу отмеченной связи (8.21) между производными и конечными разностями, очень грубо можно заменить M_{n+1} на величину $\dfrac{\max\limits_i \left\{ \left|\Delta^{n+1} y_i\right| \right\}}{h^{n+1}}$, если, конечно, разности $\Delta^{n+1} y_i$ еще несут какую-то информацию об $f(x)$, т.е. если степень n интерполяционного многочлена берется заниженной по сравнению с той, какой она должна быть для достижения максимальной точности при конкретном порядке практически постоянных разностей.

Это достигается применением неравенства

$$|q(q-1)\ldots(q-n)| \le \frac{n!}{4} \qquad \forall\, q \in (0, 1) \quad \forall\, n \in \mathbf{N}$$

(читателю предлагается доказать его самостоятельно методом математической индукции).

Подстановка последнего неравенства в (8.37) приводит к простой точечной оценке

$$|f(\tilde{x}) - P_n(x_0 + \tilde{q}h)| \le \frac{M_{n+1}}{4\,(n+1)} h^{n+1} \quad (\text{при } \tilde{q} \in (0, 1)), \qquad (8.38)$$

подчеркивающей степенную зависимость точности интерполирования от малости шага таблицы. Если шаг h мал и с ростом n не происходит слишком быстрого роста $\left|f^{(n+1)}(x)\right|$ при $x \in (a, b)$, то, как видно из (8.38), повышение степени n интерполяционного многочлена $P_n(x_0 + qh)$ за счет множителя h^{n+1} влечет уменьшение погрешности интерполяции (здесь имеется в виду только методическая погрешность, т.е. не учитывается точность задания значений функции и шум округлений). В этом случае можно говорить о точечной сходимости процесса конечноразностного интерполирования.

С равномерной сходимостью дело обстоит сложнее. Существуют примеры бесконечно дифференцируемых функций, для которых максимальная погрешность при интерполировании на заданном отрезке по системе равноотстоящих узлов не стремится к нулю при $n \to \infty$ $(h \to 0)$. Таким примером служит функция $y = \dfrac{1}{1 + 25x^2}$, впервые рассмотренная с этой целью Рунге*). Установлено [3, 134, 142 и др.], что при ее интерполировании на отрезке $[-1, 1]$ по системе равноотстоящих узлов с шагом $h = \dfrac{2}{n}$ многочленами $P_n(x)$ имеет место

$$\lim_{n\to\infty} \max_{x \in [-1, 1]} |y - P_n(x)| = \infty.$$

Завершая разговор о конечноразностной интерполяции, отметим, что наряду с уже построенными здесь интерполяционными многочленами, можно строить ряд других. Для этого существует некая схема, называемая ***диаграммой Фрезера*** [19 и др.],

*) Рунге Карл Давид Тольме (1856–1927) — немецкий физик и математик.

включающая в себя известные интерполяционные многочлены и служащая источником получения новых, предполагающих использование зигзагообразных путей продвижения по таблице конечных разностей.

Замечание 8.1. Формулы Стирлинга (8.32) и Бесселя (8.33) лишь условно можно назвать интерполяционными. Это связано со способом их получения — суммированием с коэффициентами 0.5 действительно интерполяционных формул Гаусса. Легко видеть, что, например, в линейном случае получаемая в результате такого суммирования многочленов $\overline{P}_1(x_0+qh)=y_0+q\Delta y_0$ и $\widetilde{P}_1(x_0+qh)=y_0+q\Delta y_{-1}$ формула Стирлинга

$$f(x)\approx y_0+q\frac{\Delta y_{-1}+\Delta y_0}{2}=y_0+\frac{y_1-y_{-1}}{2h}(x-x_0)=:P_s^1$$

определяется значениями данной функции $y=f(x)$ в трех точках x_{-1}, x_0 и x_1 (отсюда априори ее более высокая точность), но условие интерполяции выполняется лишь в одной из них, x_0 (рис. 8.2). Забегая вперед, скажем, что $\frac{y_1-y_{-1}}{2h}$ лучше аппроксимирует $f'(x)$, чем $\frac{y_1-y_0}{h}$ или $\frac{y_0-y_{-1}}{h}$ (сравните (13.18) с (13.14) и (13.15)), т.е. формула Стирлинга ближе к формуле Тейлора, чем формула Гаусса или Ньютона, а значит многочлен Стирлинга лучше аппроксимирует функцию $f(x)$ в окрестности точки x_0.

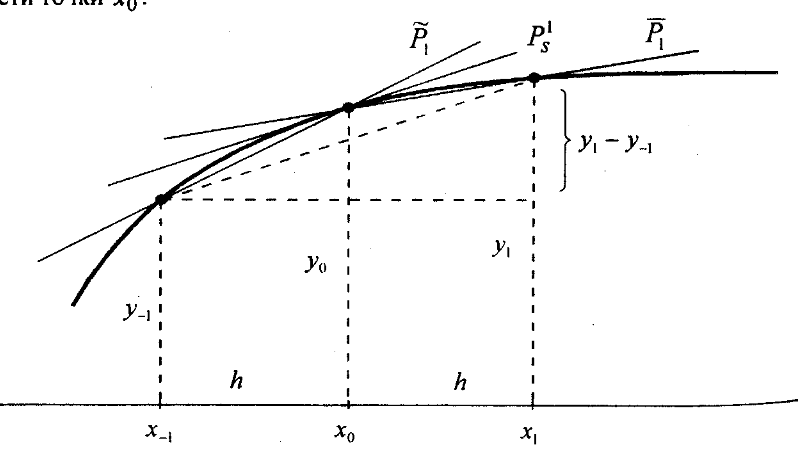


Рис. 8.2. Геометрическая интерпретация многочлена Стирлинга первой степени.

Формула Бесселя нулевого порядка

$$f(x)\approx\frac{y_0+y_1}{2}$$

определяется значениями функции $y=f(x)$ в узлах x_0 и x_1 и не является точной ни в одном из них; следовательно, многочлен Бесселя нулевой степени $P_B^0:=\frac{y_0+y_1}{2}$ не удовлетворяет условиям интерполяции (8.3) и называть его интерполяционным можно лишь в некотором более широком смысле.

В связи с отмеченной особенностью формул Стирлинга и Бесселя, выделяющей их из множества других, эквивалентных интерполяционной формуле Лагранжа конечноразностных интерполяционных формул, для них должны применяться отличные от (8.35) формулы остаточных членов; их можно найти, например, в [19, 123].

8.6. ИНТЕРПОЛЯЦИОННАЯ ФОРМУЛА НЬЮТОНА ДЛЯ НЕРАВНООТСТОЯЩИХ УЗЛОВ

Для построения интерполяционных формул, имеющих перед классической интерполяционной формулой Лагранжа (8.6) преимущества, какими обладают конечноразностные формулы, и применимых в более общем по сравнению с последними случае произвольного расположения упорядоченных несовпадающих узлов $x_0, x_1, \ldots, x_n$ на промежутке $[a,b]$, вместо конечных разностей используют разделенные разности, или иначе, разностные отношения.

Через значения функции $f(x_0), f(x_1), \ldots, f(x_n)$ сначала определяют ***разделенные разности первого порядка***:

$$f(x_0;x_1):=\frac{f(x_1)-f(x_0)}{x_1-x_0},$$

$$f(x_1;x_2):=\frac{f(x_2)-f(x_1)}{x_2-x_1},$$

$$\ldots\ldots\ldots\ldots$$

$$f(x_{n-1};x_n):=\frac{f(x_n)-f(x_{n-1})}{x_n-x_{n-1}}.$$

На этих разностях базируются ***разделенные разности второго порядка***:

$$f(x_0;x_1;x_2):=\frac{f(x_1;x_2)-f(x_0;x_1)}{x_2-x_0},$$

$$\ldots\ldots\ldots\ldots$$

$$f(x_{n-2};x_{n-1};x_n):=\frac{f(x_{n-1};x_n)-f(x_{n-2};x_{n-1})}{x_n-x_{n-2}},$$

и т.д. Таким образом, если определены k*-е разностные отношения* $f(x_i; x_{i+1}; \ldots; x_{i+k})$, то $(k+1)$-е определяются через них равенством

$$f(x_{i-1}; x_i; \ldots; x_{i+k}) = \\ = \frac{f(x_i; x_{i+1}; \ldots; x_{i+k}) - f(x_{i-1}; x_i; \ldots; x_{i+k-1})}{x_{i+k} - x_{i-1}}. \quad (8.39)$$

Легко проверить, что *операция взятия разделенной разности*, как и в случае конечной разности, *аддитивна и однородна, т.е. линейна*. Кроме того, *разделенная разность есть симметрическая функция своих аргументов*, что позволяет у разделенной разности менять аргументы местами.

Доказательство симметричности опирается на **представление k-й разделенной разности через значения функции в узлах**, имеющее вид

$$f(x_i; x_{i+1}; \ldots; x_{i+k}) = \\ = \sum_{j=0}^{k} \frac{f(x_{i+j})}{(x_{i+j} - x_i)\ldots(x_{i+j} - x_{i+j-1})(x_{i+j} - x_{i+j+1})\ldots(x_{i+j} - x_{i+k})}. \quad (8.40)$$

При $k=1$ справедливость выражения (8.40) очевидна:

$$f(x_i; x_{i+1}) = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} = \frac{f(x_{i+1})}{x_{i+1} - x_i} + \frac{f(x_i)}{x_i - x_{i+1}}.$$

Для произвольных натуральных k равенство (8.40) доказывается на основе (8.39) по индукции [19].

Важно отметить, что как и для конечных разностей, взятие разделенной разности м н о г о ч л е н а понижает на единицу его степень. Это следует из рассмотрения разделенной разности $f(x_i; x_{i+1})$ степенной функции $f(x) = x^n$:

$$f(x_i; x_{i+1}) = \frac{x_{i+1}^n - x_i^n}{x_{i+1} - x_i} = x_{i+1}^{n-1} + x_{i+1}^{n-2} \cdot x_i + \ldots + x_{i+1} \cdot x_i^{n-2} + x_i^{n-1}.$$

Отсюда приходим к заключению, что *разделенные разности n-го порядка многочлена n-й степени*

$$P_n(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1} x + a_n$$

постоянны *).

Следовательно, как и в конечноразностном случае, анализируя таблицу разделенных разностей (табл.8.8), по порядку

*) Показано [19], что $P_n(x; x+h_1; \ldots; x+h_n) \equiv a_0$, а $(n+1)$-е и все последующие разделенные разности равны нулю.

Таблица 8.8

Таблица разделенных разностей

x_i	$f(x_i)$	$f(x_i;x_{i+1})$	$f(x_i;x_{i+1};x_{i+2})$	$f(x_i;x_{i+1};x_{i+2};x_{i+3})$	$f(x_i;x_{i+1};x_{i+2};x_{i+3};x_{i+4})$	...
x_0	$f(x_0)$	$f(x_0;x_1)$	$f(x_0;x_1;x_2)$	$f(x_0;x_1;x_2;x_3)$	$f(x_0;x_1;x_2;x_3;x_4)$	...
x_1	$f(x_1)$	$f(x_1;x_2)$	$f(x_1;x_2;x_3)$	$f(x_1;x_2;x_3;x_4)$	$f(x_1;x_2;x_3;x_4;x_5)$	...
x_2	$f(x_2)$	$f(x_2;x_3)$	$f(x_2;x_3;x_4)$	$f(x_2;x_3;x_4;x_5)$	$f(x_2;x_3;x_4;x_5;x_6)$	...
x_3	$f(x_3)$	$f(x_3;x_4)$	$f(x_3;x_4;x_5)$	$f(x_3;x_4;x_5;x_6)$	...	
x_4	$f(x_4)$	$f(x_4;x_5)$	$f(x_4;x_5;x_6)$	...		
x_5	$f(x_5)$	$f(x_5;x_6)$	...			
x_6	$f(x_6)$	...				
...	...					

почти совпадающих разделенных разностей можно сделать вывод о предпочтительной степени многочлена, подходящего для интерполирования данной функции.

В основу построения интерполяционного многочлена по типу первого интерполяционного многочлена Ньютона (8.25), но для случая неравных промежутков между узлами x_i, положим следующие рассуждения.

Пусть $\varphi(x)$ — некоторая функция с известными значениями в узлах $x_0, x_1, \dots,$ а x — произвольная фиксированная точка. По определению разделенной разности первого порядка имеем

$$\varphi(x;\, x_0) = \frac{\varphi(x_0) - \varphi(x)}{x_0 - x} = \frac{\varphi(x) - \varphi(x_0)}{x - x_0},$$

откуда

$$\varphi(x) = \varphi(x_0) + \varphi(x;\, x_0)(x - x_0). \tag{8.41}$$

Для разделенной разности второго порядка по точкам $x,\ x_0,\ x_1$ записываем представление

$$\varphi(x;\, x_0;\, x_1) = \frac{\varphi(x_0;\, x_1) - \varphi(x;\, x_0)}{x_1 - x} = \frac{\varphi(x;\, x_0) - \varphi(x_0;\, x_1)}{x - x_1},$$

следствием которого является выражение

$$\varphi(x;\, x_0) = \varphi(x_0;\, x_1) + \varphi(x;\, x_0;\, x_1)(x - x_1).$$

Подставляя его в (8.41), приходим к равенству

$$\varphi(x) = \varphi(x_0) + \varphi(x_0;\, x_1)(x - x_0) + \varphi(x;\, x_0;\, x_1)(x - x_0)(x - x_1).$$

Формально, на основе определяющего разделенные разности рекуррентного соотношения (8.39), этот процесс может быть продолжен. В результате можно записать формулу, описывающую своеобразное разложение $\varphi(x)$ по произведениям разностей $(x - x_i)$, коэффициентами в котором являются разделенные разности различных порядков:

$$\begin{aligned}\varphi(x) &= \varphi(x_0) + \varphi(x_0;\, x_1)(x - x_0) + \varphi(x_0;\, x_1;\, x_2)(x - x_0)(x - x_1) + \dots \\ &\dots + \varphi(x_0;\, x_1;\, \dots;\, x_n)(x - x_0)\dots(x - x_{n-1}) + \\ &+ \varphi(x;\, x_0;\, x_1;\, \dots;\, x_n)(x - x_0)\dots(x - x_{n-1})(x - x_n).\end{aligned} \tag{8.42}$$

Если $\varphi(x) \equiv P_n(x)$ — многочлен степени n, то процесс подобного разложения исчерпывается. Разложение будет состоять из $n+1$ слагаемого, и все они будут иметь конкретные коэффициенты, так как последняя, содержащая x, разделенная разность в (8.42), т.е. $\varphi(x;\, x_0;\, \dots;\, x_n) \equiv P_n(x;\, x_0;\, \dots;\, x_n)$ имеет $(n+1)$-й порядок и, значит, равна нулю. Таким образом, для любого многочлена степени n справедливо тождество

$$\begin{aligned}P_n(x) &= P_n(x_0) + P_n(x_0;\, x_1)(x - x_0) + \\ &+ P_n(x_0;\, x_1;\, x_2)(x - x_0)(x - x_1) + \dots \\ &\dots + P_n(x_0;\, x_1;\, \dots;\, x_n)(x - x_0)(x - x_1)\dots(x - x_{n-1}).\end{aligned}$$

Предположим, что этот многочлен $P_n(x)$ является интерполяционным для некоторой функции $f(x)$. Тогда во всех узлах $x_0, x_1, \dots, x_n$ он должен иметь одинаковые с ней значения, а значит должны быть одинаковыми и их разделенные разности. Отсюда приходим к ***интерполяционной формуле Ньютона для неравноотстоящих узлов***:

$$\begin{aligned}f(x) \approx P_n(x) &:= f(x_0) + f(x_0;\, x_1)(x - x_0) + \\ &+ f(x_0;\, x_1;\, x_2)(x - x_0)(x - x_1) + \dots \\ &\dots + f(x_0;\, x_1;\, \dots;\, x_n)(x - x_0)(x - x_1)\dots(x - x_{n-1}).\end{aligned} \tag{8.43}$$

Подставив $f(x)$ вместо $\varphi(x)$ в (8.42), с учетом (8.43) получаем точное равенство

$$f(x) = P_n(x) + f(x;\, x_0;\, \dots;\, x_n)(x - x_0)\dots(x - x_{n-1})(x - x_n),$$

второе слагаемое в котором может рассматриваться в качестве остаточного члена, т.е.

$$R_n(x) := f(x) - P_n(x) = f(x;\, x_0;\, \dots;\, x_n)\Pi_{n+1}(x), \tag{8.44}$$

где $\Pi_{n+1}(x)$ — многочлен, введенный ранее посредством (8.10).

Поскольку для вычисления разности $f(x;\, x_0;\, \dots;\, x_n)$ требуется знание значения $f(x)$ наряду с известными значениями $f(x_0), \dots, f(x_n)$, представляемое формулой (8.44) выражение $R_n(x)$ фактически можно использовать только для оценивания погрешности интерполирования по формуле (8.43) через максимальные величины модулей разделенных разностей $(n+1)$-го порядка (если в них содержится достаточно незашумленной информации) или для получения других выражений остаточного члена при тех или иных предположениях о данной функции. В частности, если функция $f(x)$ имеет $(n+1)$-ю производную, то остаточный член (8.44) может быть преобразован к виду (8.12), согласно утверждению о единственности интерполяционного многочлена Лагранжа, одной из форм представления которого является (8.43).

Таблица 8.9

Разделенные разности функции $f(x)=\cos(e^{-x})$

x_i	$f(x_i)$	$f(x_i;x_{i+1})$	$f(x_i;\ldots;x_{i+2})$	$f(x_i;\ldots;x_{i+3})$	$f(x_i;\ldots;x_{i+4})$	$f(x_i;\ldots;x_{i+5})$	$f(x_i;\ldots;x_{i+6})$	$f(x_i;\ldots;x_{i+7})$
0.0	0.5403	0.7140	-0.5046	0.2124	-0.0576	0.0097	-0.0006	-0.0002
0.2	0.6831	0.4617	-0.3134	0.1317	-0.0382	0.0080	-0.0012	
0.5	0.8216	0.2423	-0.1554	0.0629	-0.0182	0.0036		
0.9	0.9185	0.1024	-0.0610	0.0229	-0.0055			
1.4	0.9697	0.0353	-0.0197	0.0060				
2.0	0.9909	0.0097	-0.0041					
2.7	0.9977	0.0016						
4.0	0.9998							

При практическом использовании интерполяционной формулы (8.43) приходится полагаться на убывание модулей слагаемых в $P_n(x)$ при увеличении номера слагаемого. Такое убывание обычно происходит до некоторых пор; затем начинается рост их модулей из-за влияния ошибок округления.

Пример 8.5. Проследим процесс формирования результата интерполирования функции $f(x)=\cos(e^{-x})$ в точке $x=0.3$ с помощью нескольких ее приближенных значений, заданных на неравномерной сетке. Эти заданные значения и вычисленные по ним разделенные разности всех возможных порядков представлены в табл.8.9.

Полагая в (8.43) $x=0.3$, $x_0=0.2$ и, далее, выбирая из табл.8.9 нужные для применения (8.43) данные, получаем:

$$f(0.3)\approx 0.6831+0.4617\cdot 0.1-0.3134\cdot 0.1\cdot(-0.2)+0.1317\cdot 0.1\cdot(-0.2)\cdot(-0.6)-$$
$$-0.0382\cdot 0.1\cdot(-0.2)\cdot(-0.6)\cdot(-1.1)+0.0080\cdot 0.1\cdot(-0.2)\cdot(-0.6)\cdot(-1.1)\cdot(-1.7)\approx$$
$$\approx 0.6831+0.04617+0.00627+0.00158+0.00050+0.00018\approx 0.7378$$

Найденное значение отличается от истинного значения $\cos(e^{-0.3})$ не более, чем на 0.0001, что в условиях точности исходных данных в пределах 0.00005 можно считать вполне хорошим результатом. Обратим внимание на предсказанное постепенное убывание роли последующих слагаемых в его формировании.

Заметим, что при использовании разделенных разностей, как и при конечноразностной интерполяции, можно менять последовательность подключения узлов и получать интерполяционные формулы для неравных промежутков, отличные от (8.43), но имеющие такую же структуру.

8.7. ОБРАТНОЕ ИНТЕРПОЛИРОВАНИЕ

Под ***задачей обратного интерполирования*** понимается задача нахождения приближенного значения $\hat{x}$ аргумента x таблично заданной функции $f(x)$, соответствующего некоторому известному значению $\hat{y}$ этой функции. При этом, естественно, предполагается, что $\hat{y}$ не совпадает ни с одним из данных значений $y_i=f(x_i)$ (и значит, $\hat{x}$ не совпадает ни с одним из узлов x_i) и, в то же время, достаточно хорошо «вписывается» в таблицу значений y_i.

Для того, чтобы говорить о теоретически однозначной разрешимости задачи обратного интерполирования, нужно потребовать выполнения основного условия существования обратной функции $x=f^{-1}(y)$ — монотонность данной сеточной функции.

Условимся далее считать, что такая монотонность имеет место либо во всей исходной таблице значений y_i, либо в некоторой ее части (используемой для решения обратной задачи).

Так как функция $y = f(x)$, вообще говоря, не может быть восстановлена точно по нескольким своим значениям (за определенными исключениями), то и обратная задача заведомо может решаться лишь приближенно.

Формально простейший из приемов решения задачи обратной интерполяции заключается в перемене ролями функции и аргумента и применения интерполяционной формулы Лагранжа, т.е. непосредственное вычисление $\hat{x}$ по формуле (сравните с (8.6))

$$\hat{x} \approx \sum_{i=0}^{n} \frac{(\hat{y}-y_0)(\hat{y}-y_1)\ldots(\hat{y}-y_{i-1})(\hat{y}-y_{i+1})\ldots(\hat{y}-y_n)}{(y_i-y_0)(y_i-y_1)\ldots(y_i-y_{i-1})(y_i-y_{i+1})\ldots(y_i-y_n)} \cdot x_i. \quad (8.45)$$

Такой подход обладает характерными для применений формулы Лагранжа недостатками (обсужденными ранее, см., в частности, § 8.3) и привлекателен лишь в случаях, когда степень n интерполяционного многочлена невысока и заранее известна.

Например, им можно воспользоваться для следующего вывода ***метода обратной линейной интерполяции*** (иначе, ***метода хорд***) решения скалярных уравнений вида $f(x) = 0$ (см. гл. 5).

Пусть для непрерывной функции $y = f(x)$ известны две точки x_0, x_1 такие, что $f(x_0) \cdot f(x_1) < 0$, т.е. между ними имеется точка $\hat{x}$ такая, что $\hat{y} := f(\hat{x}) = 0$. Приближенно эту точку можно найти обратной линейной интерполяцией, подставив в (8.45) $n = 1$ и $\hat{y} = 0$:

$$\hat{x} \approx \frac{0-y_1}{y_0-y_1} x_0 + \frac{0-y_0}{y_1-y_0} x_1.$$

Обозначая правую часть этого приближенного равенства через x_2 и учитывая, что $y_0 = f(x_0)$, $y_1 = f(x_1)$, полученному придаем вид

$$\hat{x} \approx x_2 := \frac{x_0 f(x_1) - x_1 f(x_0)}{f(x_1) - f(x_0)}.$$

На основе последней формулы (возможно, более узнаваемой в виде $x_2 = x_1 - \frac{f(x_1)(x_1 - x_0)}{f(x_1) - f(x_0)}$) строится итерационный процесс получения последовательных приближений x_k к корню $\hat{x}$ уравнения $f(x) = 0$ (см. рис.8.3, а также рис. 5.4 и формулу 5.4).

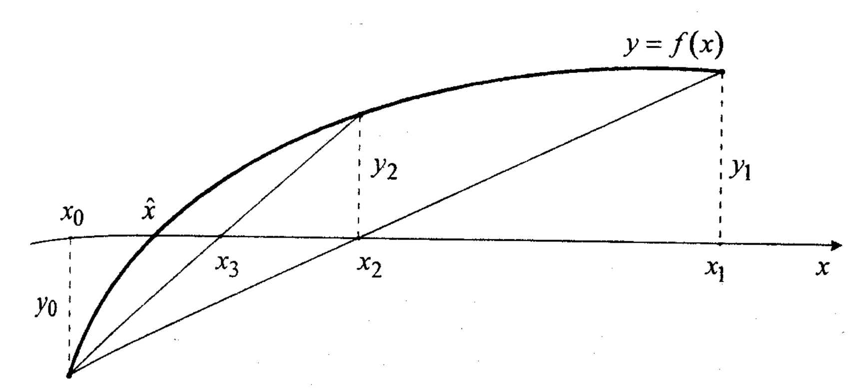


Рис. 8.3. Приближения к корню $\hat{x}$ уравнения $f(x) = 0$ методом обратной линейной интерполяции

Используя сразу большее, чем в предыдущем случае, количество точечной информации о данной функции $f(x)$, т.е. подменяя функцию $f(x)$ интерполяционным многочленом $P_n(x)$ достаточно высокой степени, можно за один шаг найти удовлетворительное приближение к искомому значению $\hat{x} = f^{-1}(\hat{y})$ и, в частности, при $\hat{y} = 0$ с хорошей точностью найти корень уравнения $f(x) = 0$.

Если сопрягать точность получения искомого значения $\hat{x} = f^{-1}(\hat{y})$ с построением соответствующего интерполяционного многочлена подходящей (неизвестной заранее) степени, то для этой цели более пригодны интерполяционные формулы, в которых можно постепенно наращивать точность за счет последовательного подключения все новых узлов. Рассмотрим, как это может быть реализовано в случае, когда информация об $f(x)$ берется на системе равноотстоящих узлов.

Предположим, что число $\hat{y}$ близко к $y_0 = f(x_0)$, например, вписывается между y_0 и $y_1 = f(x_1)$, и пусть известно достаточное количество значений y_i функции $f(x)$ в точках $x_i = x_0 + ih$ при $i = 0, 1, 2, \ldots$. Тогда за основу может быть принята первая интерполяционная формула Ньютона (8.26) с базовым узлом x_0. В силу выдвинутого ранее требования о монотонности $f(x)$, искомая точка $\hat{x}$ должна быть близка к x_0; следовательно, приближенное значение $\hat{x}$ будет найдено, если удастся найти такое (вообще говоря, малое, укладывающееся в границы интервала $(0, 1)$) значение $\hat{q} = \frac{\hat{x} - x_0}{h}$, которое является приближенным корнем

уравнения

$$y_0 + q\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_0 + \ldots + \frac{q(q-1)\ldots(q-n+1)}{n!}\Delta^n y_0 = \hat{y}. \quad (8.46)$$

Учитывая специфику уравнения (8.46), придадим ему вид задачи о неподвижной точке (6.3)

$$q = \frac{\hat{y} - y_0}{\Delta y_0} - \frac{\Delta^2 y_0}{2!\Delta y_0} q(q-1) - \ldots - \frac{\Delta^n y_0}{n!\Delta y_0} q(q-1)\ldots(q-n+1).$$

К последнему можно применить метод простых итераций (6.1), т.е. находить последовательные приближения q_k к $\hat{q}$ по формуле

$$q_{k+1} = \frac{\hat{y} - y_0}{\Delta y_0} - \frac{\Delta^2 y_0}{2!\Delta y_0} q_k(q_k - 1) - \ldots$$

$$\ldots - \frac{\Delta^n y_0}{n!\Delta y_0} q_k(q_k-1)\ldots(q_k - n + 1) \quad (8.47)$$

при $k = 0, 1, 2, \ldots$, начиная с $q_0 = \frac{\hat{y} - y_0}{\Delta y_0}$. Степень n интерполяционного многочлена здесь фиксируется в соответствии с поведением конечных разностей (не обязательно сразу максимальная; можно ее наращивать постепенно от итерации к итерации), а число итераций k, при котором следует положить $\hat{q} :\approx q_k$ определяется практическим совпадением q_k с q_{k-1} в пределах тех знаков, на которые можно рассчитывать при той или иной точности и разреженности исходной дискретной информации об $f(x)$. После того, как при некотором $k \geq 1$ будет принято $\hat{q} :\approx q_k$, считаем $\hat{x} :\approx x_0 + \hat{q}h$.

Для решения задачи обратной интерполяции можно составить и другие аналогичные (8.46) уравнения с помощью подходящих для того или иного случая разностных интерполяционных формул. Например, на базе интерполяционной формулы Ньютона для неравноотстоящих промежутков (8.43) можно построить итерационный процесс вычисления последовательных приближений $x^{(k)}$ к искомому значению $\hat{x}$ по формуле

$$x^{(k+1)} = \varphi\left(x^{(k)}\right); \quad k = 0, 1, 2, \ldots; \quad x^{(0)} = x_0,$$

где

$$\varphi(x) := x_0 - \frac{1}{f(x_0;x_1)}\left[f(x_0) - \hat{y} + f(x_0;x_1;x_2)(x-x_0)(x-x_1) + \ldots\right.$$

$$\left.\ldots + f(x_0;\, x_1;\ldots;\, x_n)(x-x_0)(x-x_1)\ldots(x-x_{n-1})\right].$$

Процесс подключения слагаемых в правой части, содержащих разделенные разности все более высоких порядков, здесь также может быть организован последовательным.

Пример 8.6. Пусть известно несколько приближенных значений функции $f(x) = \frac{\ln(x\ln(1+x))}{x}$, и по ним построена таблица конечных разностей (табл.8.10). Требуется найти приближенно корень уравнения $f(x) = 0$.

Таблица 8.10

Таблица конечных разностей функции $y = \frac{\ln(x\ln(1+x))}{x}$

x_i	y_i	Δy_i	$\Delta^2 y_i$	$\Delta^3 y_i$	$\Delta^4 y_i$
1.2	–0.046				
		0.107			
1.3	0.061		–0.023		
		0.084		0.006	
1.4	0.145		–0.017		–0.003
		0.067		0.003	
1.5	0.212		–0.014		
		0.053			
1.6	0.265				

Полагая $\hat{y} = 0$, $x_0 = 1.2$ и $n = 3$, в соответствии с формулой последовательных приближений (8.47) и данными табл.8.10, находим:

$$q_0 = \frac{0 - (-0.046)}{0.107} \approx 0.430;$$

$$q_1 = 0.430 - \frac{(-0.023)}{2\cdot 0.107}\cdot 0.430\cdot(0.430-1) - \frac{0.006}{6\cdot 0.107}\cdot 0.430\cdot(0.430-1)(0.430-2) \approx$$

$$\approx 0.430 - 0.026 - 0.004 = 0.400;$$

$$q_2 = 0.430 - \frac{(-0.023)}{2\cdot 0.107}\cdot 0.400\cdot(0.400-1) - \frac{0.006}{6\cdot 0.107}\cdot 0.400\cdot(0.400-1)(0.400-2) \approx$$

$$\approx 0.430 - 0.026 - 0.004 = 0.400.$$

Совпадение q_2 с q_1 в третьем знаке после запятой говорит о том, что можно принять $\hat{q} \approx q_2 \approx 0.400$. Учитывая, что $h = 0.1$, в силу связи $\hat{x} = x_0 + qh$, корень уравнения $\frac{\ln(x\ln(1+x))}{x} = 0$ считаем равным

$$\hat{x} \approx 1.2 + 0.400\cdot 0.1 = 1.240.$$

8.8. ИНТЕРПОЛЯЦИЯ С КРАТНЫМИ УЗЛАМИ

Рассмотрим задачу полиномиальной интерполяции функции $y = f(x)$ в более общей постановке.

Пусть на промежутке $[a, b] \subseteq D(f)$ расположены $m+1$ несовпадающих узлов $x_0, x_1, \ldots, x_m$, и пусть в этих точках известны значения $y_0 = f(x_0)$, $y_1 = f(x_1), \ldots$, $y_m = f(x_m)$ данной функции, а также некоторые ее производные (максимальный порядок производных в разных узлах различен; в каких-то узлах производные могут быть вовсе неизвестны). Такие узлы будем называть ***кратными узлами***. Конкретнее, будем считать, что заданы:

$$\begin{aligned} &\text{в узле } x_0 \quad \text{значения } y_0, y_0', \ldots, y_0^{(k_0-1)}, \\ &\text{в узле } x_1 \quad \text{значения } y_1, y_1', \ldots, y_1^{(k_1-1)}, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &\text{в узле } x_m \quad \text{значения } y_m, y_m', \ldots, y_m^{(k_m-1)}; \end{aligned} \qquad (8.48)$$

тогда ***кратность*** узла x_0 считается равной k_0, узла x_1 — k_1, ..., узла x_m — k_m.

Предполагая, что *суммарная кратность узлов* есть

$$k_0 + k_1 + \ldots + k_m = n+1, \qquad (8.49)$$

ставим задачу построения многочлена $H_n(x)$ степени n (не выше n) такого, что

$$H_n^{(j)}(x_i) = y_i^{(j)} \quad \forall\, i \in \{0, 1, \ldots, m\} \quad \forall\, j \in \{0, 1, \ldots, k_i - 1\}, \qquad (8.50)$$

где $m \geq 0$, $y_i^{(j)} := f^{(j)}(x_i)$ — заданные посредством (8.48) значения функции $f(x)$ и ее производных и по определению считается $H_n^{(0)}(x_i) := H_n(x_i)$, $y_i^{(0)} := y_i$. Многочлен $H_n(x)$ будем называть ***интерполяционным многочленом Эрмита****), а совокупность требований (8.50) — ***условиями эрмитовой интерполяции***.

Формально можно считать, что нахождение такого многочлена состоит в том, чтобы однозначно определить $n+1$ коэффициентов $a_0, a_1, \ldots, a_n$ его канонического представления

$$H_n(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_{n-1} x + a_n \qquad (8.51)$$

*) Эрми́т Шарль (1822–1901) — французский математик. Принятое обозначение многочлена $H_n(x)$ связано с французским написанием фамилии Hermite.

из условий (8.50). В силу предположения (8.49) о суммарной кратности узлов эрмитовой интерполяции, совокупность требований (8.50) можно рассматривать как систему из $n+1$ уравнений относительно $n+1$ неизвестных — коэффициентов a_k многочлена (8.51):

$$\begin{array}{llll} H_n(x_0) = y_0; & H_n'(x_0) = y_0'; & \cdots & H_n^{(k_0-1)}(x_0) = y_0^{(k_0-1)}; \\ H_n(x_1) = y_1; & H_n'(x_1) = y_1'; & \cdots & H_n^{(k_1-1)}(x_1) = y_1^{(k_1-1)}; \\ \ldots & \ldots & \ldots & \ldots \\ H_n(x_m) = y_m; & H_n'(x_m) = y_m'; & \cdots & H_n^{(k_m-1)}(x_m) = y_m^{(k_m-1)}. \end{array}$$

Единственность многочлена $H_n(x)$, удовлетворяющего условиям эрмитовой интерполяции, доказывается, как обычно, от противного. А именно, если $H_n(x)$ и $\widetilde{H}_n(x)$ — два многочлена степени n, удовлетворяющие одним и тем же $n+1$ условиям (8.50), то это означает, что многочлен-разность $H_n(x) - \widetilde{H}_n(x)$ степени не выше n имеет $n+1$ корень (с учетом кратностей), следовательно, $H_n(x) - \widetilde{H}_n(x) \equiv 0$, отсюда — совпадение $H_n(x)$ и $\widetilde{H}_n(x)$.

Существование многочлена $H_n(x)$, удовлетворяющего требованиям (8.50), можно вывести из его единственности. Действительно, возьмем в качестве исходной функции $y = f(x)$ функцию $y = 0$. Все ее значения и значения производных равны нулю, поэтому условия (8.50) для нее имеют вид

$$H_n^{(j)}(x_i) = 0, \qquad (8.52)$$

и таких условий $n+1$. Получается, что многочлен n-й степени $H_n(x)$ имеет $n+1$ корень (с учетом кратностей); значит, это есть нуль-многочлен. А это означает, что все a_k в его выражении (8.51) равны нулю. С другой стороны, (8.52) — это однородная система линейных алгебраических уравнений относительно коэффициентов a_k, которая, в силу единственности такого набора a_k, имеет т о л ь к о тривиальное решение. Отсюда следует неравенство нулю ее детерминанта, влекущее разрешимость такой системы при любых правых частях, т.е. формальное существование наборов коэффициентов a_k ($k = 0, 1, \ldots, n$), обеспечивающих выполнение условий эрмитовой интерполяции (8.50) для л ю б о й заданной с помощью (8.48) функции $y = f(x)$.

Выявление общего вида интерполяционных многочленов Эрмита $H_n(x)$ представляет непростую задачу и требует привлечения определенных сведений из теории функций комплексной переменной [129 и др.]. Рассмотрим одну из возможных процедур фактического построения таких многочленов, не требующую знания их общего вида (см., например, [19]; другой способ можно найти в [158]).

Пусть $L_m(x)$ — интерполяционный многочлен Лагранжа, построенный по данным $m+1$ значениям $y_i = f(x_i)$, $i = 0, 1, \ldots, m$. Как и ранее (см.(8.10)), будем пользоваться обозначением $\Pi_{m+1}(x) := (x - x_0)(x - x_1)\ldots(x - x_m)$. Так как по условию m заведомо не превосходит n, то по теореме о делении многочлена с остатком искомый многочлен Эрмита $H_n(x)$ можно представить в виде

$$H_n(x) = L_m(x) + H_{n-(m+1)} \cdot \Pi_{m+1}(x), \qquad (8.53)$$

где $H_{n-(m+1)}(x)$ — некоторый неизвестный пока многочлен степени $n - m - 1$.

Для построения многочлена $H_{n-(m+1)}(x)$ будем привлекать информацию о производных данной функции, т.е. равенства $H'_n(x_i) = y'_i$ в тех узлах x_i, где первые производные, в соответствии с (8.48), заданы (информация о самих значениях функции уже полностью исчерпана: в силу $L_m(x_i) = y_i$ и $\Pi_{m+1}(x_i) = 0$ для всех x_i от x_0 до x_m, согласно (8.53), будет и $H_n(x_i) = y_i$ при любых $i \in \{0, 1, \ldots, m\}$).

Продифференцировав равенство (8.53), имеем

$$H'_n(x) = L'_m(x) + H'_{n-(m+1)}(x) \cdot \Pi_{m+1}(x) + H_{n-(m+1)}(x) \cdot \Pi'_{m+1}(x). \qquad (8.54)$$

Поскольку $\Pi_{m+1}(x_i) = 0$, в тех узлах x_i, где по условию эрмитовой интерполяции справедливо $H'_n(x_i) = y'_i$, можно записать

$$L'_m(x_i) + H_{n-(m+1)}(x_i) \cdot \Pi'_{m+1}(x) = y'_i.$$

Отсюда выражаем значения многочлена $H_{n-(m+1)}(x)$ в этих узлах:

$$H_{n-(m+1)}(x_i) = \frac{y'_i - L'_m(x_i)}{\Pi'_{m+1}(x_i)}.$$

Правая часть этого равенства может быть вычислена; обозначим ее через z'_i. Таким образом, в ряде узлов x_i известны значения многочлена $H_{n-(m+1)}(x_i) = z'_i$, по которым этот многочлен

однозначно восстанавливается обычной лагранжевой интерполяцией, если в условиях (8.48) не содержится производных порядка, выше первого (т.е. нет ни одного узла кратности больше 1); подстановка найденного многочлена $H_{n-(m+1)}(x)$ в (8.53) приводит к искомому интерполяционному многочлену Эрмита. Если же в исходной информации (8.48) об $f(x)$ имеются значения производных более высокого порядка, чем первый, то для восстановления многочлена $H_{n-(m+1)}(x)$ ставится задача эрмитовой же интерполяции, для чего наряду с полученными его значениями z'_i, находят значения его производных путем дифференцирования равенства (8.54) (возможно неоднократного, в зависимости от максимального порядка заданных производных функции $f(x)$). Эта процедура построения интерполяционных многочленов Эрмита все более низких степеней продолжается до исчерпывания всей информации (8.48) о функции и ее производных.

Рассмотрим реализацию описанного процесса эрмитовой интерполяции на простом примере, демонстрирующем возможность восстановления многочлена n-й степени по его значениям и значениям некоторых его производных при суммарной кратности узлов $n+1$.

Пример 8.7. Пусть сведения о некоторой функции $y = f(x)$ представлены следующей дискретной информацией:

i	x_i	y_i	y'_i	y''_i
0	-1	0	-2	
1	0	1	0	-4
2	1	0	2	

В соответствии с обозначениями (8.48) здесь: $m = 2$; $k_0 - 1 = 1$, $k_1 - 1 = 2$, $k_2 - 1 = 1 \Rightarrow n + 1 = k_0 + k_1 + k_2 = 7 \Rightarrow n = 6$. Таким образом, по данным сведениям о функции $y = f(x)$, сосредоточенным в трех узлах $x_0 = -1$, $x_1 = 0$, $x_2 = 1$ кратности, соответственно, 2, 3, 2, следует строить интерполяционный многочлен Эрмита $H_6(x)$.

Согласно предложенной выше схеме, сначала, пользуясь столбцами x_i, y_i таблицы данных, записываем интерполяционный многочлен Лагранжа второй степени

$$L_2(x) = \frac{x(x-1)}{(-1)(-2)} \cdot 0 + \frac{(x+1)(x-1)}{1 \cdot (-1)} \cdot 1 + \frac{(x+1)x}{2 \cdot 1} \cdot 0 = 1 - x^2.$$

Далее по формуле (8.53) представляем $H_6(x)$ через $L_2(x)$, $H_3(x)$ и $\Pi_3(x)$:

$$H_6(x) = L_2(x) + H_3(x) \cdot \Pi_3(x) = 1 - x^2 + H_3(x)(x+1)x(x-1), \qquad (8.55)$$

и дифференцируем этот многочлен дважды:

$$H_6'(x) = -2x + \left(3x^2 - 1\right)H_3(x) + (x+1)x(x-1)H_3'(x);$$

$$H_6''(x) = -2 + 6x \cdot H_3(x) + 2\left(3x^2 - 1\right)H_3'(x) + (x+1)x(x-1)H_3''(x).$$

Подстановкой в $H_6'(x)$ значений $x = -1$, $x = 0$ и $x = 1$ и приравниванием $H_6'(x_i)$ заданным значениям y_i' получаем значения $H_3(x_i)$:

при $x = -1$ $\quad H_6'(-1) = 2 + 2H_3(-1) = -2 \quad \Rightarrow \quad H_3(-1) = -2;$

при $x = 0$ $\quad H_6'(0) = 0 - H_3(0) = 0 \quad \Rightarrow \quad H_3(0) = 0;$

при $x = 1$ $\quad H_6'(1) = -2 + 2H_3(1) = 2 \quad \Rightarrow \quad H_3(1) = 2.$

Учитывая их, из условия $H_6''(0) = -4$, т. е. из $-2 - 2H_3'(0) = -4$, находим $H_3'(0) = 1$.

Итак, для выявления многочлена $H_3(x)$ в (8.55) снова имеем задачу эрмитовой интерполяции с данными, содержащимися в следующей таблице:

i	x_i	$H_3(x_i)$	$H_3'(x_i)$
0	–1	–2	
1	0	0	1
2	1	2	

Здесь: $m = 2$; $k_0 = 1$, $k_1 = 2$, $k_2 = 1$ $\Rightarrow$ $n = 3$ $\Rightarrow$ $n - (m+1) = 0$. В соответствии с (8.53), для этого случая записываем:

$$H_3(x) = \widetilde{L}_2(x) + H_0 \cdot \Pi_3(x) = \frac{x(x-1)}{(-1)(-2)}(-2) + \frac{(x+1)(x-1)}{1 \cdot (-1)} \cdot 0 + \frac{(x+1)x}{2 \cdot 1} \cdot 2 +$$

$$+ H_0 \cdot (x+1)x(x-1) = 2x + x\left(x^2 - 1\right)H_0$$

(где $\widetilde{L}_2(x)$ — многочлен Лагранжа, интерполирующий функцию $H_3(x)$).

Остается найти постоянную H_0, для чего воспользуемся условием $H_3'(0) = 1$. Имеем:

$$H_3'(x) = 2 + \left(3x^2 - 1\right)H_0 \quad \Rightarrow \quad 2 + (3 \cdot 0 - 1)H_0 = 1 \quad \Rightarrow \quad H_0 = 1.$$

Следовательно,

$$H_3(x) = 2x + x\left(x^2 - 1\right) \cdot 1 = x^3 + x.$$

Подставив это в (8.55), получаем окончательное выражение искомого многочлена:

$$H_6(x) = 1 - x^2 + \left(x^3 + x\right)\left(x^3 - x\right) = x^6 - 2x^2 + 1.$$

Легко убедиться, что исходная таблица была составлена именно для этой функции.

В заключение отметим, что для $(n+1)$-кратно дифференцируемой функции $f(x)$ ***остаточный член интерполяционной формулы Эрмита*** имеет вид [19, 123, 169]

$$f(x) - H_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{k_0}(x - x_1)^{k_1} \ldots (x - x_m)^{k_m}, \qquad (8.56)$$

где $k_0 + k_1 + \ldots + k_m = n + 1$ — суммарная кратность узлов $x_0, x_1, \ldots, x_m$, а ξ — некоторая точка из промежутка $[a, b] \supseteq [x_0, x_m]$. Характерно, что в случае, когда все узлы — простые (однократные), т.е. $m = n$, интерполяционный многочлен Эрмита есть не что иное, как обычный многочлен Лагранжа $L_n(x)$, и остаточный член (8.56) совпадает с выведенным ранее его остаточным членом (8.12). Если же вся информация об $f(x)$ сосредоточена в одном узле x_0, т.е. x_0 является узлом кратности $n+1$, то многочлен Эрмита — это просто многочлен Тейлора с остаточным членом $\dfrac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1}$.

УПРАЖНЕНИЯ*)

8.1. Для функции $y = f(x)$, заданной тремя значениями $f(1) = 0.71$, $f(2) = 3.31$ и $f(3) = 0.18$, найдите коэффициенты интерполирующего ее многочлена $P_2(x) = a_0 + a_1 x + a_2 x^2$ непосредственно из условий интерполяции.

8.2. Привлекая интерполяционные соображения, докажите, что выражение

$$\frac{(x-b)(x-c)}{(a-b)(a-c)} + \frac{(x-a)(x-c)}{(b-a)(b-c)} + \frac{(x-a)(x-b)}{(c-a)(c-b)}$$

тождественно равно 1 при любых попарно различающихся a, b и c.

8.3. Дана таблица значений функции $y = \lg x$:

x_i	11	12	13	14	15
y_i	1.0414	1.0792	1.1139	1.1461	1.1761

А) С помощью линейной интерполяционной формулы Лагранжа вычислите $\lg 11.6$ $(\approx L_1(11.6))$, оцените погрешность и сравните ее с фактической ошибкой.

---

*) Предполагается, что все фигурирующие в упражнениях функции обладают нужной гладкостью.

Б) С какой точностью можно вычислить по этим данным $\lg 11.6$ посредством интерполяционной формулы Лагранжа третьей степени? Запишите расчетную формулу для вычисления $\lg 11.6 \approx L_3(11.6)$.

В) Можно ли в данных условиях построить интерполяционный многочлен пятой степени?

8.4. Пользуясь интерполяционной схемой Эйткена, пополните таблицу

x	2.0	2.1	2.2	2.3	2.4	2.5
y	0.3010		0.3424		0.3802	0.3979

недостающими значениями $y(2.1)$ и $y(2.3)$.

8.5. Какую точность можно гарантировать при линейной интерполяции функции $y = xe^{-x}$ на отрезке $[0, 1]$, считая узлами интерполяции его концы?

8.6. Восстановите многочлен $P_3(x)$ по его значениям:

x	−1	0	1	2
$P_3(x)$	1	1	−1	7

8.7. По данным упражнения 8.3 постройте таблицу конечных разностей. Вычислите по интерполяционным формулам Ньютона приближенные значения $\lg 11.6$, $\lg 10.5$, $\lg 14.5$ и $\lg 15.2$. Сравните их с точными значениями.

8.8. Применяя наиболее подходящие центральные интерполяционные формулы, найдите значения $y(1)$, $y(1.12)$ и $y(1.5)$, если соответствие между x и $y = y(x)$ задано таблицей

x	0.1	0.5	0.9	1.3	1.7	2.1
y	0.000	0.100	0.354	0.734	1.189	1.676

8.9. В каких точках выполняется условие лагранжевой интерполяции:

а) для многочлена Бесселя второй степени, построенного по точкам x_{-1}, x_0, x_1, x_2 ?

б) для многочлена Стирлинга третьей степени, построенного по точкам x_{-2}, x_{-1}, x_0, x_1, x_2 ?

8.10. По аналогии с формулой (8.43) выведите вторую интерполяционную формулу Ньютона, пригодную для интерполирования по неравным промежуткам в конце таблицы значений сеточной функции.

8.11. Для функции $y = f(x)$, заданной таблицей

x	0.5	0.7	1.0	1.4	2.0	2.6	4.0
$f(x)$	−0.555	−0.239	0.000	0.114	0.139	0.123	0.082

составьте таблицу разделенных разностей, запишите подходящие для вычисления $f(0.6)$, $f(1.5)$ и $f(3)$ конкретные интерполяционные многочлены и найдите эти приближенные значения.

8.12. Докажите, что в примере 8.6 (см. § 8.7) выполняются достаточные условия сходимости итерационного процесса (8.47).

8.13. Некоторая зависимость $y = y(x)$ задана следующей таблицей:

x	1	7	13	19	25
y	0.21361	0.40541	0.85225	1.79725	3.78183

Обратным интерполированием установите, каким значениям x отвечают значения $y = 0.3178$ и $y = 1.159$.

8.14. Найдите многочлен $P(x)$, данные о котором представлены следующей таблицей:

x	$P(x)$	$P'(x)$	$P''(x)$
−1	15	−14	−2
0	4	−7	
2	18		

ГЛАВА 9 **МНОГОЧЛЕНЫ ЧЕБЫШЕВА И НАИЛУЧШИЕ РАВНОМЕРНЫЕ ПРИБЛИЖЕНИЯ**

Дается определение многочленов Чебышева первого рода и изучаются их свойства. Показывается, что погрешность интерполирования гладкой функции многочленом фиксированной степени будет наименьшей, когда в качестве узлов интерполяции используются корни многочленов Чебышева. Приводится (без доказательства) теорема о чебышевском альтернансе, служащая теоретической основой построения наилучших равномерных приближений, и рассматриваются простейшие ситуации, когда такие многочлены могут быть построены. Обсуждается идея использования разложения функций в степенные ряды для получения многочленов, близких к многочленам наилучших равномерных приближений.

9.1. ОПРЕДЕЛЕНИЕ И СВОЙСТВА МНОГОЧЛЕНОВ ЧЕБЫШЕВА

Определение 9.1. ***Многочленом Чебышева****) *называется функция*

$$T_n(x) := \cos(n \arccos x), \qquad (9.1)$$

где $n \in \mathbf{N}_0$, $x \in [-1,1]$.

Прежде всего, убедимся, что функция $T_n(x)$, представленная с помощью тригонометрических функций, на самом деле является многочленом при любом $n = 0, 1, 2, \ldots$.

Непосредственной подстановкой в (9.1) значений $n=0$ и $n=1$ получаем $T_0(x)=1$, $T_1(x)=x$.

Положив $\alpha := \arccos x$, имеем:

$$T_1(x) = \cos\alpha, \quad T_n(x) = \cos n\alpha,$$

$$T_{n-1}(x) = \cos(n-1)\alpha, \quad T_{n+1}(x) = \cos(n+1)\alpha,$$

*) Чебышёв Пафнутий Львович (1821–1894) — знаменитый русский математик. Его работы о многочленах наилучшего равномерного приближения легли в основу конструктивной теории функций. Стандартное обозначение $T_n(x)$ можно соотнести с немецким написанием фамилии Tschebyschew.

и так как (по формуле суммы косинусов)

$$\cos(n+1)\alpha + \cos(n-1)\alpha = 2\cos\alpha\cos n\alpha,$$

то, значит, справедливо равенство

$$T_{n+1}(x) + T_{n-1}(x) = 2T_1(x)T_n(x),$$

которое может быть переписано в виде

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \qquad (9.2)$$

Формула (9.2) определяет при $n = 1, 2, 3, \ldots$ последовательность функций $T_n(x)$, начинающуюся с $T_0(x)=1$, $T_1(x)=x$, рекуррентно; при этом нужно иметь в виду, что здесь $x \in [-1, 1]$, как и в (9.1).

Подставляя в (9.2) заданные начальные члены последовательности $(T_n(x))$, найдем несколько ее последующих членов:

$$T_2(x) = 2x^2 - 1;$$

$$T_3(x) = 2x(2x^2-1) - x = 4x^3 - 3x;$$

$$T_4(x) = 2x(4x^3-3x) - 2x^2 + 1 = 8x^4 - 8x^2 + 1;$$

$$T_5(x) = 2x(8x^4-8x^2+1) - 4x^3 + 3x = 16x^5 - 20x^3 + 5x$$

и т. д.

Графики нескольких многочленов Чебышева (с первого по четвертый) изображены на рис. 9.1.

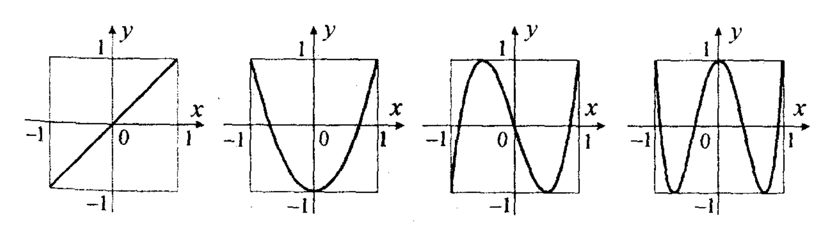


Рис. 9.1. Графики многочленов $T_1(x)$, $T_2(x)$, $T_3(x)$ *и* $T_4(x)$

Анализ рекуррентной формулы (9.2) позволяет считать очевидными следующие факты:

1) *все функции* $T_n(x)$, *определенные в* (9.1), *являются многочленами при любом натуральном* n;

2) *степени этих многочленов возрастают с увеличением* n, *причем старший член многочлена* $T_n(x)$ *равен* $2^{n-1}x^n$;

3) *многочлены $T_n(x)$ при четных n выражаются через степенные функции только четных степеней, при нечетных — только нечетных.*

Наряду с многочленами Чебышева $T_n(x)$, часто используют многочлены, получаемые из $T_n(x)$ делением на старший коэффициент, т.е.

$$\hat{T}_n(x) := \frac{1}{2^{n-1}} T_n(x)$$

— многочлены со старшим коэффициентом 1. Будем называть их ***нормированными многочленами Чебышева***.

Многочлены Чебышева обладают рядом замечательных свойств. Рассмотрим некоторые их **свойства**, имеющие отношение к поставленной выше проблеме аппроксимации функций.

Свойство 9.1. *Многочлен Чебышева $T_n(x)$ (а значит, и многочлен $\hat{T}_n(x)$) имеет на отрезке $[-1, 1]$ ровно n различных действительных корней; все они задаются формулой*

$$x_k = \cos\frac{2k+1}{2n}\pi \text{ , где } k = 0, 1, \ldots, n-1. \quad (9.3)$$

Доказательство. Корни многочлена Чебышева $T_n(x)$ можно получить, решая тригонометрическое уравнение

$$\cos(n \arccos x) = 0.$$

Имеем:

$$n \arccos x = \frac{\pi}{2} + \pi k \quad \Leftrightarrow \quad \arccos x = \frac{2k+1}{2n}\pi.$$

Поскольку главные значения арккосинуса должны принадлежать промежутку $[0, \pi]$, целая переменная k здесь должна принимать значения от 0 до $n-1$. Переходя на этом промежутке к обратной функции, иначе, беря косинус от левой и правой частей последнего равенства и учитывая, что $\cos\arccos x = x$, приходим к выводу, что n действительных корней многочлена $T_n(x)$, определяемые формулой (9.3), на самом деле существуют. Их несовпадение и принадлежность отрезку $[-1, 1]$ следует из того, что они суть значения косинуса, соответствующие различным значениям его аргумента, расположенным на промежутке его монотонности.

Свойство 9.2. *Корни многочленов Чебышева перемежаются с точками их наибольших и наименьших значений, равных соответственно $+1$ и -1 для $T_n(x)$ и $+\frac{1}{2^{n-1}}$ и $-\frac{1}{2^{n-1}}$ для $\hat{T}_n(x)$. А именно, при $j = 0, 1, \ldots, n$ имеют место экстремумы*

$$T_n(x_j) = (-1)^j, \quad \hat{T}_n(x_j) = \frac{(-1)^j}{2^{n-1}} \text{ в точках } x_j = \cos\frac{j}{n}\pi. \quad (9.4)$$

Доказательство. Из определения многочлена $T_n(x)$ формулой (9.1) следует, что $|T_n(x)| \le 1$. Покажем, что максимальное значение 1 для $|T_n(x)|$ достигается, т.е. существуют такие точки $x \in [-1, 1]$, в которых $|T_n(x)| = 1$. Имеем:

$$T_n(x) = \pm 1 \quad \Leftrightarrow \quad \cos(n \arccos x) = \pm 1 \quad \Leftrightarrow$$

$$n \arccos x = \pi j \ (j \in Z) \quad \Leftrightarrow \quad \arccos x = \frac{j}{n}\pi \ (j \in Z).$$

Число $\frac{j}{n}\pi$ принадлежит отрезку $[0, \pi]$, т.е. может служить значением арккосинуса, если $j \in \{0, 1, \ldots, n\}$. Следовательно, уравнение $|T_n(x)| = 1$ имеет своими корнями $n+1$ точек

$$x_j = \cos\frac{j}{n}\pi, \quad j = 0, 1, \ldots, n. \quad (9.5)$$

Что эти точки экстремумов многочлена $T_n(x)$ перемежаются с его корнями, легко увидеть, переписав (9.5) в виде $x_j = \cos\frac{2j}{2n}\pi$ и сравнив это с (9.3). Факт чередования максимумов и минимумов у $T_n(x)$ следует из того, что значения выражения $n \arccos x$ в точках x_j, согласно промежуточному равенству $n \arccos x = \pi j$, равны поочередно $0, \pi, 2\pi, 3\pi, \ldots$, и потому, в зависимости от четности j, значения $\cos(n \arccos x)$ будут соответственно то $+1$, то -1.

Свойство 9.3 (*теорема Чебышева*). *Из всех многочленов степени n со старшим коэффициентом 1 нормированный многочлен Чебышева $\hat{T}_n(x)$ наименее уклоняется от нуля на отрезке $[-1, 1]$.*

Доказательство (от противного). Пусть существует многочлен $\widetilde{P}_n(x) = x^n + \widetilde{a}_1 x^{n-1} + \ldots + \widetilde{a}_n$ такой, что

$$\max_{x\in[-1,1]} \left|\widetilde{P}_n(x)\right| < \max_{x\in[-1,1]} \left|\hat{T}_n(x)\right| \quad \left(= \frac{1}{2^{n-1}} \text{ по свойству 9.2}\right).$$

Разность этих двух многочленов, очевидно, есть многочлен степени, не выше $n-1$; обозначим его

$$Q_{n-1}(x) := \hat{T}_n(x) - \widetilde{P}_n(x) = b_1 x^{n-1} + b_2 x^{n-2} + \ldots + b_n.$$

Пользуясь тем, что график $\hat{T}_n(x)$ лежит в полосе, ограниченной прямыми $y = \pm\frac{1}{2^{n-1}}$, чередуя касания то верхней, то нижней из этих прямых (по свойству 9.2), а график $\widetilde{P}_n(x)$ при $x \in [-1, 1]$ должен лежать строго внутри этой полосы (по предположению), то можно утверждать, что многочлен-разность $Q_{n-1}(x)$ в точках экстремумов $x_j = \cos\frac{j}{n}\pi$ должен иметь определенные знаки: «+» при четных j и «–» при нечетных j (рис. 9.2).

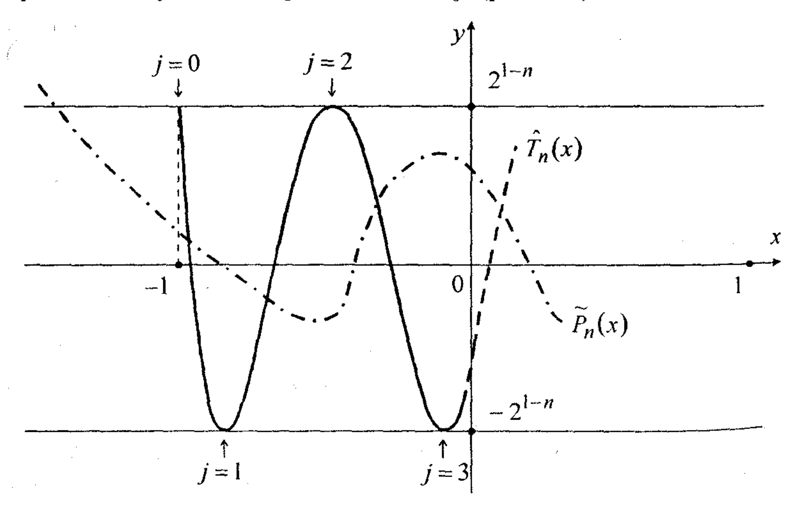


Рис. 9.2. Возможное совместное поведение графиков $\hat{T}_n(x)$ и $\widetilde{P}_n(x)$

Поскольку $\hat{T}_n(x)$ на $[-1, 1]$ имеет $n+1$ таких точек экстремумов, следовательно, многочлен $Q_{n-1}(x)$ должен иметь, по меньшей мере, n перемен знаков, т.е. n корней, что противоречит следствию из основной теоремы алгебры многочленов. Полученное противоречие говорит о несостоятельности сделанного в начале доказательства предположения.

Только что доказанное свойство означает, что *среди всех многочленов степени n вида*

$$P_n(x) = x^n + a_1 x^{n-1} + \ldots + a_n \tag{9.6}$$

именно нормированный многочлен Чебышева $\hat{T}_n(x)$ минимизирует максимальное расстояние от графика многочлена при $x \in [-1, 1]$ до оси абсцисс, т.е. $\hat{T}_n(x)$ — это многочлен с наименьшей нормой (так называемой чебышевской нормой) *на множестве многочленов вида* (9.6) *в пространстве* $C[-1, 1]$.

Замечание 9.1. Иногда бывает более удобным использовать ***смещенные многочлены Чебышева*** $T_n^*(x)$, которые определяются на отрезке $[0, 1]$ и могут быть получены из классических многочленов Чебышева $T_n(x)$ заменой в них аргумента x на аргумент $2x - 1$.

9.2. ИНТЕРПОЛЯЦИЯ ПО ЧЕБЫШЕВСКИМ УЗЛАМ

Вернемся к изучавшейся в предыдущей главе задаче интерполяции.

Сравнивая конечноразностные интерполяционные многочлены, построенные по системе равноотстоящих узлов, с интерполяционным многочленом Лагранжа, предполагающим произвольное расположение несовпадающих узлов на промежутке интерполирования $[a, b]$, следует отметить, что первые более просты и удобны в использовании, вторые же обладают бóльшими возможностями. Нет сомнений в том, что если можно располагать узлы в пределах отрезка $[a, b]$ как угодно, то имеет смысл использовать большее количество точечной информации о функции там, где она более сильно изменяется. Особенно существенным это замечание может оказаться при эрмитовой интерполяции.

Подойдем к проблеме расположения узлов интерполяции с несколько иной стороны.

Желая, чтобы интерполяционный многочлен Лагранжа $L_n(x)$ (1.6) в целом хорошо приближал функцию $y = f(x)$ на отрезке $[a, b]$, поставим вопрос: как расположить на нем $n+1$

узлов интерполяции x_i $(i=0, 1, \ldots, n)$, чтобы при этом минимизировать максимальную на $[a, b]$ погрешность?

Поскольку, согласно (1.15),

$$\max_{x\in[a,b]}|f(x)-L_n(x)|\le\frac{M_{n+1}}{(n+1)!}\max_{x\in[a,b]}|\Pi_{n+1}(x)|, \tag{9.7}$$

и величина $M_{n+1}:=\max_{x\in[a,b]}\left|f^{(n+1)}(x)\right|$ не зависит от расположения узлов, займемся изучением определенного в (1.10) многочлена $\Pi_{n+1}(x)$.

Сначала выполним одно простое преобразование независимой переменной, которое неоднократно будет использоваться и в дальнейшем. А именно, положим

$$x=\frac{a+b}{2}+\frac{b-a}{2}t. \tag{9.8}$$

Легко видеть, что эта линейная замена осуществляет взаимнооднозначное соответствие между $x\in[a, b]$ и $t\in[-1, 1]$.

Благодаря (9.8) узлам $x_i\in[a, b]$ можно сопоставить точки $t_i\in[-1, 1]$, полагая

$$x_i=\frac{a+b}{2}+\frac{b-a}{2}t_i.$$

Тогда образующие $\Pi_{n+1}(x)$ сомножители $x-x_i$ преобразуются по формуле

$$x-x_i=\frac{b-a}{2}(t-t_i),$$

и, следовательно,

$$\Pi_{n+1}(x):=\prod_{i=0}^{n}(x-x_i)=\left(\frac{b-a}{2}\right)^{n+1}\cdot\prod_{i=0}^{n}(t-t_i)=$$

$$=\left(\frac{b-a}{2}\right)^{n+1}\cdot(t-t_0)(t-t_1)\ldots(t-t_n). \tag{9.9}$$

Из (9.9) явствует, что значение $\max_{x\in[a,b]}|\Pi_{n+1}(x)|$ будет минимальным, когда будет минимальным значение $\max_{t\in[-1,1]}|\Pi_{n+1}(t)|$, где

$$\Pi_{n+1}(t):=(t-t_0)(t-t_1)\ldots(t-t_n)$$

— многочлен $n+1$ степени, в котором точки $t_i\in[-1, 1]$ $(i=0, 1, \ldots, n)$ считаем параметрами. Так как коэффициент при старшей степени многочлена $\Pi_{n+1}(t)$ в его каноническом представлении равен 1, то, в силу свойства 9.3, величина $\max_{t\in[-1,1]}|\Pi_{n+1}(t)|$ (а значит, и $\max_{x\in[a,b]}|\Pi_{n+1}(x)|$) будет минимальной в том случае, когда $\Pi_{n+1}(t)$ есть нормированный многочлен Чебышева $\hat{T}_{n+1}(t)$.

Таким образом, полагая $\Pi_{n+1}(t)=\hat{T}_{n+1}(t)$, имеем: с одной стороны, точки $t_0, t_1, \ldots, t_n$ по замыслу являются узлами интерполяции (для переведенной задачи интерполяции с исходного отрезка $[a, b]$ на отрезок $[-1, 1]$), с другой стороны, в силу структуры многочлена $\Pi_{n+1}(t)$, они служат его корнями, а значит, и корнями $\hat{T}_{n+1}(t)$. Отсюда — вывод:

максимальная погрешность интерполирования достаточно гладкой функции на отрезке $[-1, 1]$ многочленом n-й степени будет минимальной, когда в качестве узлов интерполяции $t_0, t_1, \ldots, t_n\in[-1, 1]$ берутся корни многочлена Чебышева $T_{n+1}(t)$ (как известно, совпадающие с корнями $\hat{T}_{n+1}(t)$). Будем называть их ***чебышевскими узлами интерполяции***.

Знание экстремальных значений многочлена Чебышева (см. (9.4)) позволяет уточнить величину максимального отклонения $L_n(x)$ от $f(x)$ при таком выборе узлов, т.е. когда точки t_i $\left(=\cos\frac{2i+1}{2n+2}\pi,\ i=\overline{0, n}\right)$ есть корни $\hat{T}_{n+1}(t)$, а $x_i=\frac{a+b}{2}+\frac{b-a}{2}t_i$.

А именно, так как $\max_{t\in[-1,1]}\left|\hat{T}_{n+1}(t)\right|=\frac{1}{2^n}$, то, согласно (9.9),

$$\max_{x\in[a,b]}|\Pi_{n+1}(t)(x)|=\left(\frac{b-a}{2}\right)^{n+1}\cdot\frac{1}{2^n}=\frac{(b-a)^{n+1}}{2^{2n+1}};$$

подставляя это в неравенство (9.7), получаем

$$\max_{x\in[a,b]}|f(x)-L_n(x)|\le\frac{M_{n+1}}{(n+1)!}\frac{(b-a)^{n+1}}{2^{2n+1}}. \tag{9.10}$$

Найденная оценка (9.10) называется ***наилучшей равномерной оценкой погрешности интерполяции***.

Покажем ее **неулучшаемость**, т.е. что существуют такие функции $f(x)$, для которых нестрогое неравенство (9.10) реализуется в виде равенства. С этой целью будем рассматривать аппроксимацию многочленом Лагранжа n-й степени, построенным

по чебышевским узлам интерполяции, функции $f(x)$, представляющей собой некоторый, вообще говоря, произвольный многочлен $n+1$-й степени

$$P_{n+1}(x) := a_0 x^{n+1} + a_1 x^n + a_2 x^{n-1} + \ldots + a_n x + a_{n+1}.$$

Так как для такой функции $f(x)$ производная $n+1$-го порядка есть $f^{(n+1)}(x) \equiv P_{n+1}^{(n+1)}(x) \equiv a_0(n+1)!$, то в неравенстве (9.10) можно считать $M_{n+1} = a_0(n+1)!$, и сама эта оценка (9.10) превращается в оценку

$$\max_{x\in[a,b]} |P_{n+1}(x) - L_n(x)| \le a_0 \frac{(b-a)^{n+1}}{2^{2n+1}}.$$

Рассматривая же фактическую разность между $P_{n+1}(x)$ и $L_n(x)$, в силу постоянства $(n+1)$-й производной, т.е. независимости $P_{n+1}(\xi)$ от положения точки ξ на (a, b), имеем:

$$P_{n+1}(x) - L_n(x) = \frac{P_{n+1}^{(n+1)}(\xi)}{(n+1)!}\Pi_{n+1}(x) = a_0\left(\frac{b-a}{2}\right)^{n+1} \cdot \hat{T}_{n+1}(t) \quad (9.11)$$

(см. (1.13) и (9.9); в последнем случае следует принять во внимание, что $\Pi_{n+1}(t) = \hat{T}_{n+1}(t)$ в соответствии со спецификой фиксирования точек t_i). Зная точное значение $\max\limits_{t\in[-1,1]} |\hat{T}_{n+1}(t)|$, на основе (9.11) получаем:

$$\max_{x\in[a,b]} |P_{n+1}(x) - L_n(x)| = a_0\left(\frac{b-a}{2}\right)^{n+1} \cdot \frac{1}{2^n} = a_0 \frac{(b-a)^{n+1}}{2^{2n+1}}.$$

Как видим, фактическая максимальная разность и ее оценка по формуле (9.10) на функции $f(x) = P_{n+1}(x)$ совпали, значит оценка (9.10) достижима, т.е. правая часть ее в общем случае не может быть уменьшена.

9.3. О МНОГОЧЛЕНАХ НАИЛУЧШИХ РАВНОМЕРНЫХ ПРИБЛИЖЕНИЙ

Резюмируя исследования предыдущего параграфа, в частности, анализируя неравенство (9.10), можно хотя бы частично ответить на вопрос о сходимости последовательности интерполяционных многочленов Лагранжа $L_n(x)$ к функции $f(x)$, для которой они построены:

если функция $f(x)$ бесконечно дифференцируема на $[a, b]$ и в качестве узлов интерполяции берутся корни многочленов Чебышева (приведенные к отрезку $[a, b]$), то

$$\max_{x\in[a,b]} |f(x) - L_n(x)| \xrightarrow[n\to\infty]{} 0.$$

Поскольку $\max\limits_{x\in[a,b]} |f(x) - L_n(x)|$ есть не что иное, как $\|f(x) - L_n(x)\|_{C[a,b]}$ — ***чебышевская норма*** в пространстве $C[a, b]$ непрерывных на $[a, b]$ функций, можно говорить, что для достаточно гладких функций $f(x)$ при специальном расположении узлов интерполяции последовательность интерполяционных многочленов Лагранжа ($L_n(x)$) (построенных по точным значениям функции $f(x)$) сходится к $f(x)$ по норме Чебышева; другими словами, имеет место р а в н о м е р н а я с х о д и м о с т ь последовательности ($L_n(x)$) к $f(x)$.

Обобщением установленного факта для непрерывных (не обязательно дифференцируемых) функций и произвольных (не обязательно интерполяционных) многочленов является широко известная в математическом анализе теорема.

Теорема 9.1 (*Вейерштрасса*) [7, 51, 82, 105, 123, 145, 173]. *Для любой непрерывной на $[a, b]$ функции $f(x)$ найдется многочлен $Q_n(x)$ такой, что*

$$|f(x) - Q_n(x)| < \varepsilon \qquad \forall\, \varepsilon > 0 \;\; \forall\, x \in [a, b].$$

Как следует из теоремы Вейерштрасса, если отказаться от того, чтобы аппроксимирующий функцию $f(x)$ многочлен $Q_n(x)$ степени n был интерполяционным, от $f(x)$ достаточно потребовать непрерывность на $[a, b]$, чтобы за счет повышения степени многочлена при надлежащем подборе его коэффициентов величина чебышевской нормы $\|f(x) - Q_n(x)\|_{C[a,b]}$ могла быть сделанной сколь угодно малой, иначе, чтобы качество аппроксимации функции $f(x)$ многочленом $Q_n(x)$ на отрезке $[a, b]$ было сколь угодно хорошим в смысле чебышевской метрики.

Если степень n аппроксимирующего $f(x)$ многочлена $Q_n(x)$ зафиксировать и распоряжаться только его коэффициентами, то в общем случае величину $\|f(x) - Q_n(x)\|_{C[a,b]}$ нельзя

сделать сколь угодно малой. Однако доказано [45], что для любой функции $f(x)\in C[a, b]$ существует единственный многочлен $Q_n^f(x)$ такой, который из всех многочленов $Q_n(x)$ степени n наилучшим образом аппроксимирует на $[a, b]$ функцию $f(x)$, минимизируя максимальное расстояние между $f(x)$ и $Q_n(x)$. Этот многочлен, т.е. многочлен $Q_n^f(x)$ такой, что

$$\left\|f(x)-Q_n^f(x)\right\|_{C[a,b]}=\inf\|f(x)-Q_n(x)\|_{C[a,b]},$$

называется ***многочленом наилучшего равномерного приближения*** для $f(x)$ на $[a, b]$ или ее ***чебышевским приближением****).

Для последовательности величин

$$E_n(f):=\left\|f(x)-Q_n^f(x)\right\|_{C[a,b]},$$

представляющих собой погрешности аппроксимации функции $f(x)\in C[a, b]$ посредством многочлена $Q_n^f(x)$, в соответствии с теоремой Вейерштрасса можно установить монотонную сходимость к нулю, т. е. что

$$E_n(f)\le E_{n-1}(f) \quad \text{и} \quad E_n(f)\xrightarrow[n\to\infty]{}0.$$

Одним из характеристических свойств многочленов наилучших равномерных приближений является **критерий Чебышева**. Его отражает следующая теорема.

Теорема 9.2 (*Чебышева*). *Многочлен $Q_n^f(x)$ является многочленом наилучшего равномерного приближения для функции $f(x)\in C[a, b]$ тогда и только тогда, когда на $[a, b]$ существует не менее $n+2$ точек x_i таких, что в них поочередно принимаются наибольшие положительные и отрицательные отклонения, т. е. поочередно разность $f(x_i)-Q_n^f(x_i)$ равна E или $-E$, где*

$$E:=\left\|f(x)-Q_n^f(x)\right\|_{C[a,b]}=\max_{x\in[a,b]}\left|f(x)-Q_n^f(x)\right|.$$

Эта теорема (доказательство которой можно найти, например, в [7, 13, 18, 173]) говорит о том, что максимальная ошибка аппроксимации функции многочленом наилучшего равномерного приближения реализуется в числе точек, большем, по меньшей мере, на 2, чем степень многочлена, причем знаки ошибки чередуются. Точки x_i, в которых реализуется максимальное отклонение многочлена $Q_n^f(x)$ от $f(x)$ на $[a, b]$, называются ***точками чебышевского альтернанса***.

*) Первое доказательство существования многочлена наилучшего равномерного приближения для произвольной непрерывной функции дано французским математиком Э. Борелем (1871–1956).

К сожалению, неизвестны ни общий вид многочленов наилучших равномерных приближений, ни способы их построения. Имеются лишь некоторые методики построения многочленов, близких к наилучшим равномерным (см., например, [18, 63, 134, 150]), а также способы построения чебышевских приближений невысокого порядка для нескольких весьма узких классов функций. Последние существенно опираются на приведенную теорему о чебышевском альтернансе, что демонстрируется в следующих двух простейших случаях.

Случай А. Пусть функция $f(x)$ непрерывна на $[a, b]$, и пусть для нее требуется построить **многочлен наилучшего равномерного приближения нулевой степени**. Обозначим этот приближающий многочлен через $\varphi_0(x)$. Он определяется всего одним параметром: $\varphi_0(x)=A_0$. Чтобы найти значение этого параметра для заданной функции $f(x)$, воспользуемся тем свойством непрерывной на замкнутом промежутке функции, согласно которому на нем всегда найдутся, по крайней мере, две точки, в которых она принимает свои наименьшее и наибольшее значения.

Пусть $\min\limits_{x\in[a,b]} f(x)=m$, $\max\limits_{x\in[a,b]} f(x)=M$. Тогда совершенно очевидно, что полагая $A_0=\dfrac{m+M}{2}$, т. е. подменяя функцию $f(x)$ функцией $\varphi_0=\dfrac{m+M}{2}$, будем иметь максимальное отклонение

$$E=\|f(x)-\varphi_0(x)\|_{C[a,b]}=\max_{x\in[a,b]}\left|f(x)-\frac{m+M}{2}\right|=$$

$$=M-\frac{m+M}{2}=\frac{m+M}{2}-m=\frac{M-m}{2},$$

причем точки отрезка $[a, b]$, в которых оно реализуется — это точки, где принимаются значения m и M. В силу непрерывности $f(x)$, локальные минимумы и максимумы должны чередоваться; по меньшей мере, два из них определяют точки чебышевского альтернанса (с чередованием знаков разностей $f(x)-\dfrac{m+M}{2}$). Поэтому не существует другой постоянной, которая приближала бы $f(x)$ на $[a, b]$ лучше, чем постоян-

ная $\frac{m+M}{2}$, в смысле чебышевской метрики (на рис. 9.3 в качестве чебышевского альтернанса могут служить пары точек x_1, x_2 или x_1, x_3).

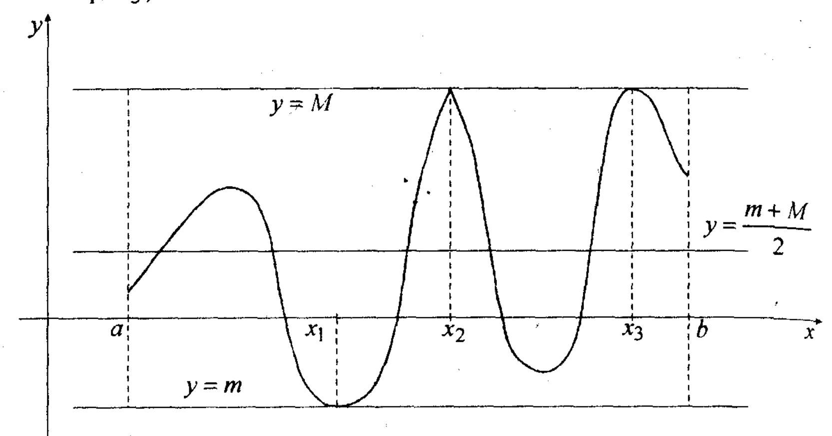


Рис. 9.3. Наилучшее равномерное приближение функции $f(x)$ с помощью постоянной

Случай Б. Пусть аппроксимируемая функция $f(x)$ дифференцируема и выпукла (в широком смысле) на отрезке $[a, b]$, а аппроксимирующая ее функция $\varphi(x) = A_0 + A_1 x$ — **многочлен наилучшего равномерного приближения первой степени**. Чтобы найти его коэффициенты A_0 и A_1, следует изучить разность между $f(x)$ и $\varphi(x)$, т. е. функцию $u(x) := f(x) - A_0 - A_1 x$.

Так как функция $f(x)$ по предположению выпукла, а сдвиг на линейную функцию $A_0 + A_1 x$ не изменяет выпуклости, то и функция $u(x)$ выпукла на $[a, b]$. Если речь идет о выпуклости вниз, то, как известно из анализа выпуклых функций, выпуклость $u(x)$ на $[a, b]$ влечет ее унимодальность на этом отрезке, т. е. существует единственная точка $c \in [a, b]$, в которой $u(x)$ имеет минимум; если $u(x)$ выпукла вверх, то в точке c должен быть максимум $u(x)$. В любом случае, точка $c \in [a, b]$ есть точка экстремума $u(x)$, и за счет возможности варьирования коэффициентов функции $\varphi(x)$ (точнее, коэффициента A_1) можно считать, что точка c является внутренней точкой отрезка $[a, b]$.

Потребуем, чтобы точки a, c и b в указанной последовательности составляли чебышевский альтернанс, т. е. чтобы в них последовательно принимались значения E, $-E$, E или $-E$, E,

$-E$, где $E := \|f(x) - \varphi(x)\|_{C[a,b]} = \max_{x \in [a,b]} |f(x) - \varphi(x)|$ — максимальная погрешность аппроксимации функции $f(x)$ функцией $\varphi(x)$. Добавляя к этим требованиям еще необходимое условие экстремума дифференцируемой функции $u(x)$ в точке c и возвращаясь к исходным функциям, приходим к системе четырех уравнений относительно четырех неизвестных A_0, A_1, E и c (из которых, в основном, лишь первые три неизвестные представляют интерес):

$$\begin{cases} f(a) - A_0 - A_1 a = E, \\ f(c) - A_0 - A_1 c = -E, \\ f(b) - A_0 - A_1 b = E, \\ f'(c) - A_1 = 0 \end{cases} \quad \text{или} \quad \begin{cases} f(a) - A_0 - A_1 a = -E, \\ f(c) - A_0 - A_1 c = E, \\ f(b) - A_0 - A_1 b = -E, \\ f'(c) - A_1 = 0. \end{cases}$$

Получить решение такой системы в общем виде не представляется возможным, поскольку неизвестная величина c входит в нее нелинейным образом (в каждом конкретном случае подобная система без проблем решается численно).

Ключом к геометрической интерпретации случая Б служит двоякое выражение коэффициента A_1 из уравнений системы: из первого и третьего уравнений, рассматриваемых совместно, имеем $A_1 = \frac{f(b) - f(a)}{b - a}$, а из четвертого следует $A_1 = f'(c)$, что говорит о параллельности хорды, стягивающей концы дуги графика $f(x)$ на $[a, b]$, касательной, проведенной к $f(x)$ во внутренней точке альтернанса, и аппроксимирующей $f(x)$ прямой $y = A_0 + A_1 x$ (рис. 9.4).

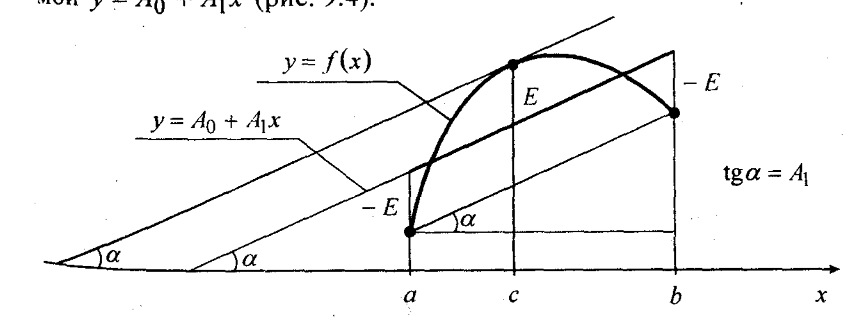


Рис. 9.4. Наилучшая линейная аппроксимация дифференцируемой выпуклой функции

Пример 9.1. Построим многочлены наилучшего равномерного приближения нулевой и первой степеней для функции $y = \sin x$ на отрезке $\left[0, \frac{\pi}{6}\right]$.

Сразу заметим, что данная функция всюду дифференцируема (а значит, непрерывна) и выпукла вверх на заданном отрезке. При этом

$$m := \min_{x \in [0, \frac{\pi}{6}]} \sin x = \sin 0 = 0, \quad M := \max_{x \in [0, \frac{\pi}{6}]} \sin x = \sin \frac{\pi}{6} = 0.5.$$

Следовательно, согласно рассмотренному выше случаю А, найдя

$$\frac{M+m}{2} = \frac{0.5+0}{2} = 0.25 \quad \text{и} \quad \frac{M-m}{2} = \frac{0.5-0}{2} = 0.25,$$

при $x \in \left[0, \frac{\pi}{6}\right]$ можно считать, что $\sin x \approx 0.25$ с предельной погрешностью 0.25.

Далее, в соответствии со случаем Б, продифференцировав данную функцию, составляем систему:

$$\begin{cases} \sin 0 - A_0 - A_1 \cdot 0 = -E, \\ \sin c - A_0 - A_1 c = E, \\ \sin \frac{\pi}{6} - A_0 - A_1 \frac{\pi}{6} = -E, \\ \cos c - A_1 = 0, \end{cases} \quad \text{т. е.} \quad \begin{cases} A_0 = E, \\ \sin c - A_0 - A_1 c = E, \\ 0.5 - A_0 - A_1 \frac{\pi}{6} = -E, \\ \cos c = A_1. \end{cases}$$

Из нее последовательно находим:

$$A_1 = 0.5 : \frac{\pi}{6} = \frac{3}{\pi} \approx 0.9549;$$

$$c = \arccos \frac{3}{\pi} \approx 0.3014;$$

$$A_0 (= E) = \frac{1}{2}(\sin c - A_1 c) \approx 0.0045.$$

Таким образом, функцию $y = \sin x$ на отрезке $\left[0, \frac{\pi}{6}\right]$ можно подменить линейной функцией $y = 0.0045 + 0.9549x$, и наибольшая ошибка при этом не будет превышать величины ≈ 0.0045.

9.4. ЭКОНОМИЗАЦИЯ СТЕПЕННЫХ РЯДОВ

Рассмотрим один из простейших способов построения многочленов, близких к наилучшим равномерным, о существовании которых упоминалось в предыдущем параграфе.

В математическом анализе хорошо изучено и широко применяется разложение функций в степенные ряды, в частности, в ряды Тейлора. Частичные суммы таких рядов — многочлены — используются в качестве локальных аппроксимаций для исходных функций. Степени используемых при этом многочленов зависят от требуемой точности аппроксимации, положения точки из области сходимости ряда, в окрестности которой производится аппроксимация, скорости сходимости ряда. В некоторых случаях такой подход мало приемлем из-за медленной сходимости рядов и большой неравномерности, т. е. существенной разницы в необходимых для заданной точности степенях приближающих многочленов при разных значениях аргумента.

Для улучшения указанных параметров частичных сумм степенных рядов можно привлечь многочлены Чебышева. *Процедура преобразования степенного ряда, представляющего собой разложение некоторой функции по системе степенных функций, в разложение ее по системе многочленов Чебышева называется* ***экономизацией степенного ряда*** [187].

Чтобы преобразовать степенной ряд в ряд по системе многочленов Чебышева, нужно сначала обратить формулы, по которым многочлены Чебышева $T_n(x)$ выражаются через степенные функции. А именно, глядя на несколько записанных в § 9.1 первых многочленов Чебышева, можно через них выразить степени x последовательно одну за другой:

$$\begin{aligned} 1 &= T_0; \\ x &= T_1; \\ x^2 &= \frac{1}{2}(T_0 + T_2); \\ x^3 &= \frac{1}{4}(3T_1 + T_3); \\ x^4 &= \frac{1}{8}(3T_0 + 4T_2 + T_4); \\ x^5 &= \frac{1}{16}(10T_1 + 5T_3 + T_5) \end{aligned} \qquad (9.12)$$

и т. д. (аргумент x в этих выражениях для краткости опущен).

Если некоторая функция $y = f(x)$ на некотором промежутке $[a, b] \subseteq [-1, 1]$ представлена степенным рядом

$$f(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_k x^k + \ldots,$$

то, подставляя сюда вместо степеней x их выражения (9.12) через многочлены Чебышева и приводя подобные члены, можно, в

принципе, получить разложение $f(x)$ вида

$$f(x) = b_0 + b_1 T_1 + b_2 T_2 + \ldots + b_n T_n + \ldots \quad (9.13)$$

Имея в виду рассмотренные в § 9.1 свойства многочленов Чебышева $T_n(x)$, можно рассчитывать, что многочлены, получаемые усечением разложений (9.13), будут близки к многочленам наилучших равномерных приближений.

Не пытаясь ответить на все возникающие здесь вопросы как теоретического, так и практического характера (некоторые сведения об этом можно найти, например, в [187]), посмотрим на примере, какой эффект может дать простейшая процедура экономизации.

Пример 9.2. Известно [59], что при $x \in (-1, 1]$ справедливо разложение

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} - \ldots \quad (9.14)$$

Ограничиваясь первыми пятью членами этого разложения, т. е. многочленом пятой степени, перейдем с помощью равенств (9.12) к приближенному представлению функции $\ln(1+x)$ через многочлены Чебышева:

$$\ln(1+x) \approx T_1 - \frac{1}{4}(T_0 + T_2) + \frac{1}{12}(3T_1 + T_3) - \frac{1}{32}(3T_0 + 4T_2 + T_4) +$$

$$+ \frac{1}{80}(10T_1 + 5T_3 + T_5) = -\frac{11}{32}T_0 + \frac{11}{8}T_1 - \frac{3}{8}T_2 + \frac{7}{48}T_3 - \frac{1}{32}T_4 + \frac{1}{80}T_5.$$

Если здесь оставить всего три первых члена, т. е. представить $\ln(1+x)$ многочленом второй степени

$$\ln(1+x) \approx -\frac{11}{32}T_0 + \frac{11}{8}T_1 - \frac{3}{8}T_2 = \frac{1}{32} + \frac{11}{8}x - \frac{3}{4}x^2,$$

и сравнить полученное квадратичное приближение функции $\ln(1+x)$ с квадратичным же тейлоровским приближением $\ln(1+x) \approx x - \dfrac{x^2}{2}$, непосредственно вытекающим из разложения (9.14), то получим картину, изображенную на рис. 9.5.

Она показывает (как это и следовало ожидать), что в середине интервала сходимости ряда (9.14) данная функция точнее аппроксимируется функцией $y = x - \dfrac{x^2}{2}$, а вблизи концов этого интервала — функцией $y = \dfrac{1}{32} + \dfrac{11}{8}x - \dfrac{3}{4}x^2$. Максимальная на $(-1, 1]$ погрешность в последнем случае меньше, но, разумеется, полученная экономизацией квадратичная функция не является многочленом наилучшего равномерного приближения среди всех многочленов второй степени.

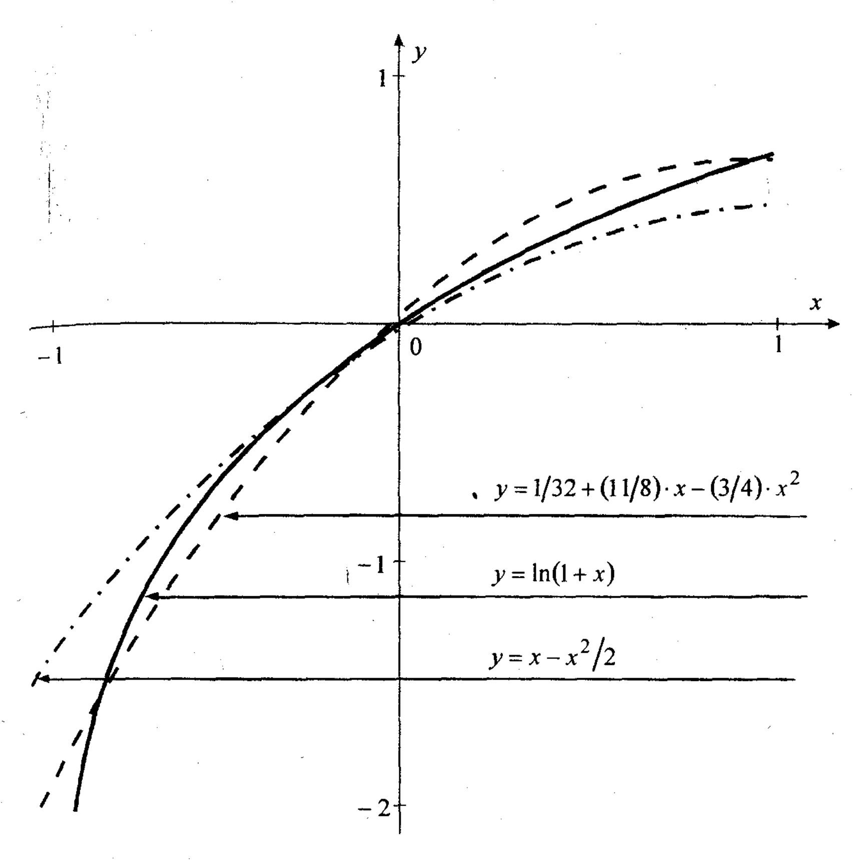


Рис. 9.5. Графики функции $\ln(1+x)$ *и двух ее квадратичных приближений*

Многочлены наилучших равномерных приближений вызывают большой теоретический и практический интерес. Например, вычисление значений некоторых основных элементарных функций на микрокалькуляторах и компьютерах базируется на том, что подсчет значения функции в произвольной точке области определения сводится к вычислению значения на некотором стандартном промежутке, на котором данная функция подменяется многочленом, близким к многочлену наилучшего равномерного приближения такой степени, при которой гарантируется, что максимальная ошибка не будет превосходить заданной фиксированной величины при любом значении аргумента из этого промежутка [61, 112].

Кроме процедуры экономизации степенных рядов, для построения многочленов, близких к многочленам наилучших равномерных приближений, привлекают и другие приемы, из которых наиболее известны алгоритмы Ремеза и Валле–Пуссена [63, 134, 145, 150].

Замечание 9.2. Можно встретить и несколько иное трактование понятия (и процедуры) ***экономизации степенных разложений***. Так, например, в [23] (см. также [44]) этим термином определяется процесс перехода от аппроксимации функции $f(x)$ многочленом n-й степени к аппроксимации многочленом $(n-1)$-й степени с сохранением (оценки) точности равномерного приближения. В основе такого процесса лежит следующее утверждение.

Теорема 9.3. *Пусть* $f(x) \approx S_n(x) := \sum_{k=0}^{n} a_k x^k$ *, причем*

$$\max_{x\in[-1,1]} |f(x) - S_n(x)| \le \delta < \varepsilon .$$

Тогда, если выполняется неравенство $\delta + \frac{|a_n|}{2^{n-1}} < \varepsilon$*, то многочлен* $P_{n-1}(x) := S_{n-1}(x) + a_n\left(x^n - \frac{1}{2^{n-1}} T_n(x)\right)$ *аппроксимирует* $f(x)$ *на отрезке* $[-1,1]$ *с оценкой*

$$\max_{x\in[-1,1]} |f(x) - P_{n-1}(x)| < \varepsilon .$$

Подробное рассомтрение этого процесса на примерах, дающее представление о его сущности, см. в книге Ланцоша [107], где вместо термина *экономизация* используется термин ***телескопический сдвиг***.

УПРАЖНЕНИЯ

9.1. Запишите несколько первых смещенных многочленов Чебышева (см. замечание 9.1). Какова формула корней смещенного многочлена Чебышева n-й степени?

9.2. Какая из функций семейства

$$\varphi(a, b, c, x) := ax^2 + bx + c$$

наиболее близка к функции $f(x) := x^3$ в том смысле, что

$$\max_{x\in[-1,1]} |f(x) - \varphi(a, b, c, x)| = \min \quad ?$$

9.3. Пусть функция $f(x)$ определена и достаточное число раз дифференцируема на отрезке $[2,4]$. В каких точках следует знать

значения $f(x)$, чтобы проинтерполировать ее с минимальной максимальной погрешностью:

а) многочленом первой степени?

б) многочленом второй степени?

9.4. Найдите неулучшаемую оценку погрешности интерполирования функции $f(x) = \ln(1+x)$ на отрезке $[0,1]$:

а) многочленом первой степени;

б) многочленом второй степени.

Постройте эти многочлены.

9.5. Постройте многочлены наилучшего равномерного приближения нулевой и первой степеней и укажите наибольшие величины погрешностей аппроксимации для функций:

а) $f(x) = \sqrt{x}$ при $x \in [0,1]$;

б) $f(x) = \frac{1}{x}$ при $x \in [1,2]$;

в) $f(x) = \ln(1+x)$ при $x \in [0,1]$;

В случае в) сравните чебышевское приближение первой степени с результатом линейной интерполяции по чебышевским узлам (см. упр. 9.4).

9.6. Используя разложение функции $f(x) = \operatorname{arctg} x$ в ряд Тейлора (Маклорена) до пятой степени x, примените процедуру экономизации для представления этой функции через многочлены Чебышева первой и третьей степеней. Дайте графическое сравнение результатов такой экономизации, аналогичное приведенному на рис. 9.5.

ГЛАВА 10 МЕТОД НАИМЕНЬШИХ КВАДРАТОВ И НАИЛУЧШИЕ СРЕДНЕКВАДРАТИЧЕСКИЕ ПРИБЛИЖЕНИЯ

Рассматриваются две простейшие задачи, приводящие к широко известному методу наименьших квадратов: решение переопределенных линейных алгебраических систем и подбор параметров эмпирических зависимостей. Далее ставится более общая задача аппроксимации функций обобщенными многочленами, минимизирующими среднеквадратические отклонения в той или иной метрике; показывается, с какими проблемами при этом можно столкнуться, если выполнять полиномиальную аппроксимацию на базе обычных степенных функций. Приводятся справочные данные о классических ортогональных многочленах, изучается простая схема построения системы ортогональных многочленов, выписываются конкретные формулы для получения с их помощью многочленов Фурье, аппроксимирующих заданные функции в интегральном или в точечном смыслах.

10.1. ПРОСТЕЙШАЯ ОБРАБОТКА ЭМПИРИЧЕСКИХ ДАННЫХ МЕТОДОМ НАИМЕНЬШИХ КВАДРАТОВ

Рассмотрим две задачи, возникающие вне математики при изучении каких-либо закономерностей на основе наблюдений или экспериментов и приводящие к истокам широко известного и весьма популярного метода наименьших квадратов*). Характерной особенностью таких задач является то обстоятельство, что исходный материал для выявления тех или иных закономерностей — заведомо приближенный, в силу неточности измерительных приборов, неповторимости условий наблюдений, случайных ошибок и пр. Отсюда естественность стремления исследователей к накоплению как можно большего количества экспериментальных данных с тем, чтобы, усредняя их какимлибо образом, получить требуемые значения, как правило, небольшого числа параметров изучаемых закономерностей наиболее точно.

*) Об этом методе и его создателях уже упоминалось вкратце в § 1.7.

Задача А. Предположим, что нужно знать значения m различных величин $x_1, x_2, \ldots, x_m$, которые нельзя измерить непосредственно, но известно, что они находятся в линейной связи, и коэффициенты этой связи могут быть приближенно получены в результате измерений. Проведя серию из n таких измерений, имеем систему линейных алгебраических уравнений (СЛАУ)

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m = b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m = b_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m = b_n. \end{cases} \qquad (10.1)$$

Чтобы эта система могла быть однозначно разрешимой, т. е. чтобы можно было найти единственный набор значений переменных $x_1, x_2, \ldots, x_m$, интересующих исследователя, число измерений n должно быть не меньшим числа неизвестных m, и при этом ранг матрицы системы должен в точности совпадать с m. В силу того, что все числа a_{ij} и b_i в (10.1) (или часть из них) являются экспериментальными данными, получаемое таким путем однозначное решение может сильно отличаться от однозначного же решения, находящегося таким же образом с помощью другой серии экспериментов или наблюдений.

Поэтому число n в системе (10.1) берут заведомо бóльшим, чем m, и рассматривают переопределенную, как правило, несовместную СЛАУ. Поскольку такая система в общем случае не имеет решения в привычном понимании, в то время как по смыслу исходной задачи решение должно существовать, следует каким-то образом обобщить понятие решения СЛАУ и получить метод его нахождения.

С этой целью будем искать набор $x_1^*, x_2^*, \ldots, x_m^*$ значений переменных $x_1, x_2, \ldots, x_m$ не такой, который обращал бы каждое из уравнений системы (10.1) в тождество, а при котором была бы минимальной сумма квадратов отклонений значений левых частей уравнений (10.1) от правых (иначе, квадратов невязок). Это означает, что вектор $\left(x_1^*; x_2^*; \ldots; x_m^*\right)$ должен быть решением задачи минимизации

$$r_1^2 + r_2^2 + \ldots + r_n^2 \to \min,$$

где $r_i := a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{im}x_m - b_i$.

Очевидно, минимизируемая функция

$$\Phi(x_1, x_2, \ldots, x_m) := \sum_{i=1}^{n} (a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{im}x_m - b_i)^2,$$

являясь неотрицательной квадратичной, имеет единственную точку минимума $\left(x_1^*; x_2^*; \ldots; x_m^*\right)$, которая может быть найдена из необходимых условий гладкого экстремума, т. е. в результате приравнивания к нулю частных производных функции $\Phi(x_1, x_2, \ldots, x_m)$, взятых по каждому из ее аргументов.

Выполняя дифференцирование, приходим к линейной же системе

$$\begin{cases} 2\sum_{i=1}^{n}(a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{im}x_m - b_i)a_{i1} = 0, \\ 2\sum_{i=1}^{n}(a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{im}x_m - b_i)a_{i2} = 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ 2\sum_{i=1}^{n}(a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{im}x_m - b_i)a_{im} = 0, \end{cases}$$

которая имеет m уравнений и m неизвестных и которой может быть придан стандартный вид СЛАУ:

$$\begin{cases} \left(\sum_{i=1}^{n}a_{i1}^2\right)x_1 + \left(\sum_{i=1}^{n}a_{i2}a_{i1}\right)x_2 + \ldots + \left(\sum_{i=1}^{n}a_{im}a_{i1}\right)x_m = \sum_{i=1}^{n}b_ia_{i1}, \\ \left(\sum_{i=1}^{n}a_{i1}a_{i2}\right)x_1 + \left(\sum_{i=1}^{n}a_{i2}^2\right)x_2 + \ldots + \left(\sum_{i=1}^{n}a_{im}a_{i2}\right)x_m = \sum_{i=1}^{n}b_ia_{i2}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \left(\sum_{i=1}^{n}a_{i1}a_{im}\right)x_1 + \left(\sum_{i=1}^{n}a_{i2}a_{im}\right)x_2 + \ldots + \left(\sum_{i=1}^{n}a_{im}^2\right)x_m = \sum_{i=1}^{n}b_ia_{im}. \end{cases} \quad (10.2)$$

Ее решение — вектор $\left(x_1^*; x_2^*; \ldots; x_m^*\right)$ — может быть получено каким-либо методом решения однозначно определенных линейных систем, в частности, учитывающим симметрию СЛАУ (10.2) (см., например, [3, 61]).

Такой способ решения переопределенных систем называется ***методом наименьших квадратов*** (МНК), а вектор $\left(x_1^*; x_2^*; \ldots; x_m^*\right)$ по отношению к исходной системе (10.1) называется ***нормальным псевдорешением*** (более подробно об этом обобщении понятия решения можно узнать, знакомясь с теорией и методами решения некорректных задач, см. § 1.7, а также [11, 28, 33, 34, 41, 93, 128, 148, 175, 194]).

Пример 10.1. Пусть определению подлежат две величины x_1 и x_2, связь между которыми задается тремя равенствами:

$$\begin{cases} x_1 + 2x_2 = 8.5, \\ 2x_1 + x_2 = 6.7, \\ x_1 + 3x_2 = 11.1, \end{cases} \quad (10.3)$$

и пусть правые части уравнений (10.3) — заведомо приближенные числа. Совершенно очевидно, что эта система противоречива, т.е. не имеет решения в обычном смысле. Для получения ее нормального псевдорешения, т.е. решения в смысле метода наименьших квадратов, составим новую систему по типу (10.2) при $m = 2$, $n = 3$:

$$\begin{cases} 6x_1 + 7x_2 = 33, \\ 7x_1 + 14x_2 = 57. \end{cases}$$

Решив последнюю, получаем пару $x_1^* = 1.8$, $x_2^* \approx 3.171$, не удовлетворяющую никакому отдельно взятому уравнению данной системы (10.3) и, в то же время, устраивающую систему в целом в оговоренном смысле. Геометрическое представление о найденном решении можно получить из рис. 10.1.

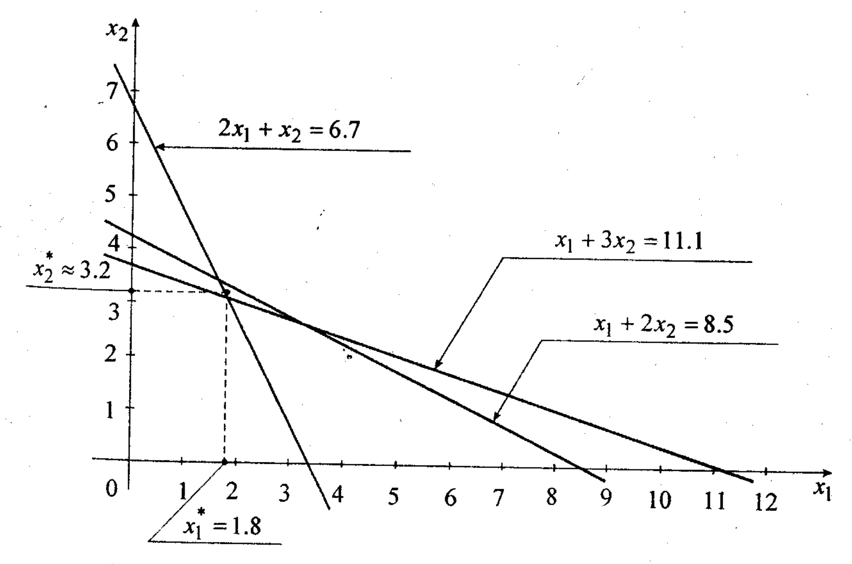


Рис. 10.1. *Геометрическая интерпретация системы* (10.3) *и ее нормального псевдорешения*

Замечание 10.1. Геометрические образы уравнений системы (10.3) не изменяются при умножении левых и правых частей уравнений на одинаковые отличные от нуля числа, т.е. можно считать, что преобразованная таким способом система в этом смысле эквивалентна исходной. Однако находимые методом наименьших квадратов нормальные псевдорешения эквивалентных в указанном смысле систем будут различаться. Например,

если второе уравнение системы (10.3) умножить на 100, т.е. заменить уравнением

$$200x_1 + 100x_2 = 670,$$

а первое и третье оставить неизменными, то система вида (10.2) для такого случая есть

$$\begin{cases} 40002x_1 + 20005x_2 = 134019.6, \\ 20005x_1 + 10013x_2 = 67050.3. \end{cases}$$

Решая ее, получаем псевдорешение $\tilde{x}_1 \approx 1.756$, $\tilde{x}_2 \approx 3.188$, близкое к найденному выше, но все же отличное от него.

Невязки r_i двумерных линейных уравнений можно интерпретировать как отклонения точки от прямых в тех случаях, когда их уравнения приводятся к нормальному виду. Деля каждое уравнение данной системы на квадратный корень из суммы квадратов его коэффициентов, приходим сначала к нормализованной системе

$$\begin{cases} \frac{1}{\sqrt{5}}x_1 + \frac{2}{\sqrt{5}}x_2 = \frac{8.5}{\sqrt{5}}, \\ \frac{2}{\sqrt{5}}x_1 + \frac{1}{\sqrt{5}}x_2 = \frac{6.7}{\sqrt{5}}, \\ \frac{1}{\sqrt{10}}x_1 + \frac{3}{\sqrt{10}}x_2 = \frac{11.1}{\sqrt{10}}, \end{cases}$$

из нее, как и выше, получаем однозначно разрешимую систему

$$\begin{cases} 1.1x_1 + 1.1x_2 = 5.49, \\ 1.1x_1 + 1.9x_2 = 8.07, \end{cases}$$

решение которой $\bar{x}_1 \approx 1.766$, $\bar{x}_2 \approx 3.225$ имеет хорошую геометрическую интерпретацию: сумма квадратов расстояний от точки $(\bar{x}_1; \bar{x}_2)$ до всех прямых, определяемых уравнениями данной системы (10.3), меньше, чем от любой другой точки $(x_1; x_2)$.

Ответ на вопрос о том, стоит ли искать такое оптимальное псевдорешение с помощью нормализации уравнений данной системы, или составлять систему вида (10.2) непосредственно из данной системы, или проводить перед ее составлением некоторое предварительное масштабирование (усиливающее или уменьшающее роль отдельных связей между искомыми величинами), в конкретных случаях зависит от содержательного смысла коэффициентов и неизвестных СЛАУ (10.1).

Задача Б. Предположим, что между независимой переменной x и зависимой переменной y имеется некая неизвестная функциональная связь $y = f(x)$. Эта связь отображается таблицей

x	x_0	x_1	...	x_n
y	y_0	y_1	...	y_n

приближенных значений $y_i \approx f(x_i)$, получаемых в ходе наблюдений или экспериментов. Требуется дать приближенное аналитическое описание этой связи, т. е. подобрать функцию $\varphi(x)$

такую, которая аппроксимировала бы на отрезке $[x_0, x_n]$ заданную отдельными приближенными значениями y_i функцию $f(x)$.

Для решения этой задачи заведомо неудачным является интерполяционный подход хотя бы потому, что функция $\varphi(x)$ такая, что $\varphi(x_i) = y_i$ при всех $i \in \{0, 1, \ldots, n\}$, будет мало похожа на искомую $f(x)$, поскольку в ней отразятся все ошибки экспериментальных данных. Уже это заставляет отказаться от идеи интерполяции и находить функцию $\varphi(x)$ такую, чтобы она хорошо отражала «в среднем» зависимость между x и y.

Конкретнее, из каких-либо соображений (аналитических, графических или иных) аппроксимирующая $f(x)$ функция $\varphi(x)$ берется из определенного m-параметрического семейства функций, и ее параметры подбираются так, чтобы сумма квадратов отклонений вычисляемых значений $\varphi(x_i)$ от заданных приближенных значений y_i была минимальной. Такая функция (т.е. при таком оптимальном наборе параметров) будет *наилучшей аппроксимацией* $f(x)$ среди функций выбранного семейства *в смысле метода наименьших квадратов*. Ясно, что число данных приближенных значений y_i в таблице должно быть не меньшим, чем число параметров в подбираемой зависимости $\varphi(x)$; как правило, считается, что $n >> m$.

Итак, согласно МНК, задаем семейство

$$y = \varphi(x, a_1, a_2, \ldots, a_m)$$

и ищем значения параметров $a_1, a_2, \ldots, a_m$ (где $m \le n-1$), решая экстремальную задачу

$$\Phi(a_1, a_2, \ldots, a_m) := \sum_{i=0}^{n} (\varphi(x_i, a_1, a_2, \ldots, a_m) - y_i)^2 \to \min.$$

Как и в задаче А, оптимальный набор параметров $a_1^*, a_2^*, \ldots, a_m^*$ может быть найден из системы

$$\begin{cases} \sum_{i=0}^{n} (\varphi(x_i, a_1, a_2, \ldots, a_m) - y_i) \left.\frac{\partial \varphi}{\partial a_1}\right|_{x = x_i} = 0, \\ \sum_{i=0}^{n} (\varphi(x_i, a_1, a_2, \ldots, a_m) - y_i) \left.\frac{\partial \varphi}{\partial a_2}\right|_{x = x_i} = 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \sum_{i=0}^{n} (\varphi(x_i, a_1, a_2, \ldots, a_m) - y_i) \left.\frac{\partial \varphi}{\partial a_m}\right|_{x = x_i} = 0, \end{cases} \qquad (10.4)$$

представляющей необходимые условия экстремума функции $\Phi(a_1, a_2, \ldots, a_m)$, в силу ее специфики, являющиеся и достаточными условиями ее минимума.

Если функция $\varphi(x, a_1, a_2, \ldots, a_m)$ есть линейная функция относительно своих параметров $a_1, a_2, \ldots, a_m$, то система (10.4) тоже будет линейной; в общем случае (10.4) — нелинейная система, что влечет за собой определенные трудности при ее решении. Спасительным в последней ситуации является тот факт, что обычно при задании семейств функций, аппроксимирующих реальные зависимости, число параметров берется небольшим (2 – 3), причем какие-то из этих параметров могут входить линейным образом.

В зависимости от характера табличных данных, изучаемого с помощью их изображения в соответствующей системе координат или с помощью некоторых прикидочных расчетов (см., например, [58, 62]), при обработке результатов экспериментов часто используют те или иные из следующих двухпараметрических семейств функций:

$$y = ax + b, \quad y = a + b\ln x \quad (y = a + b\lg x),$$
$$y = ax^b, \quad y = ae^{bx} \quad (y = a\cdot 10^{bx}),$$
$$y = a + \frac{b}{x}, \quad y = \frac{1}{ax+b}, \quad y = \frac{x}{ax+b};$$

реже применяются трехпараметрические семейства

$$y = ax^2 + bx + c, \quad y = ax^b + c,$$
$$y = ae^{bx} + c \quad (y = a\cdot 10^{bx} + c);$$

при изучении периодических явлений применяют тригонометрические функции.

В качестве примера рассмотрим ситуацию, когда есть основания считать, что значения x_i и соответствующие им приближенные значения y_i отражают некую линейную зависимость. Тогда, полагая

$$\varphi(x, a, b) = ax + b,$$

находим $\dfrac{\partial\varphi}{\partial a} = x$, $\dfrac{\partial\varphi}{\partial b} = 1$, и для вычисления параметров этой функции φ составляем систему типа (10.4):

$$\begin{cases} \sum\limits_{i=0}^{n}(ax_i + b - y_i)\cdot x_i = 0, \\ \sum\limits_{i=0}^{n}(ax_i + b - y_i)\cdot 1 = 0. \end{cases}$$

Упрощая ее, приходим к стандартной СЛАУ

$$\begin{cases} \left(\sum\limits_{i=0}^{n} x_i^2\right)a + \left(\sum\limits_{i=0}^{n} x_i\right)b = \sum\limits_{i=0}^{n} x_i y_i, \\ \left(\sum\limits_{i=0}^{n} x_i\right)a + (n+1)b = \sum\limits_{i=0}^{n} y_i \end{cases} \qquad (10.5)$$

с симметричной квадратной матрицей коэффициентов и заведомо однозначным решением $(a^*;\ b^*)$.

Заметим, что вместо того, чтобы решать нелинейные системы, получающиеся из (10.4) при поиске параметров конкретных семейств функций, когда эти параметры входят туда нелинейным образом, можно попытаться сначала л и н е а р и з о в а т ь подбираемую зависимость.

Пусть, например, в качестве аппроксимирующей $f(x)$ функции берется экспоненциальная функция

$$\varphi(x, a, b) = ae^{bx}. \qquad (10.6)$$

Найдя ее частные производные

$$\frac{\partial\varphi}{\partial a} = e^{bx}, \quad \frac{\partial\varphi}{\partial b} = abe^{bx},$$

видим, что подстановка этой функции и ее производных по параметрам в (10.4) приведет к нелинейной относительно параметра b системе. Чтобы не решать нелинейных уравнений, поступим так.

Прологарифмируем равенство

$$ae^{bx} = y,$$

получив при этом

$$\ln a + bx = \ln y,$$

введем новый параметр $A := \ln a$ и пересчитаем данную таблицу, переведя значения y_i в $Y_i := \ln y_i$. По таблице

x	x_0	x_1	…	x_n
Y	Y_0	Y_1	…	Y_n

методом наименьших квадратов находим оптимальные параметры A^* и b^* линейной функции $Y = A + bx$ (полагая $\varphi_1(x, A, b) = A + bx$, пользуемся системой (10.5), считая там $a = b$, а $b = A$). После этого вычисляем оптимальное значение $a^* = e^{A^*}$ параметра a исходной зависимости (10.6) и записываем итоговый результат

$$f(x) \approx a^* e^{b^* x}.$$

Замечание 10.2. К методу наименьших квадратов могут привести и иные соображения. Например, в курсах математической статистики к нему приходят, изучая корреляционные и функциональные зависимости при наличии случайных ошибок [30, 83, 121 и др.].

10.2. ОБОБЩЕННЫЕ МНОГОЧЛЕНЫ НАИЛУЧШИХ СРЕДНЕКВАДРАТИЧЕСКИХ ПРИБЛИЖЕНИЙ

Вернемся к общей постановке задачи аппроксимации функций. Пусть аппроксимируемая функция $f(x)$ и аппроксимирующая функция $\varphi(x)$ непрерывны на отрезке $[a,b]$ и аппроксимация должна производиться так, чтобы функция $\varphi(x)$ «в среднем хорошо описывала» поведение функции $f(x)$ при $x \in [a,b]$. Будем здесь постоянно иметь в виду две аппроксимационные ситуации: первая, это когда функция $f(x)$ считается (по крайней мере, теоретически) известной в любой точке x отрезка $[a,b]$, и близость между $f(x)$ и $\varphi(x)$ понимается в интегральном смысле, и вторая, когда $f(x)$ известна (причем приближенно) только в $n+1$ точках $x_0, x_1, \ldots, x_n$ отрезка $[a,b]$, в которых и производится согласование $f(x)$ с $\varphi(x)$ подобно тому, как это делалось при рассмотрении задачи Б в предыдущем параграфе. В связи с этим, будем параллельно рассматривать:

а) пространство $C_L[a, b]$ непрерывных на $[a, b]$ функций со скалярным произведением

$$(f, \varphi)_{C_L[a,b]} := \frac{1}{b-a}\int_a^b f(x)\varphi(x)dx, \tag{10.7}$$

метрикой (расстоянием)

$$\rho(f, \varphi)_{C_L[a,b]} := \sqrt{\frac{1}{b-a}\int_a^b (f(x)-\varphi(x))^2\,dx} \tag{10.8}$$

и нормой

$$\|f\|_{C_L[a,b]} := \rho(f, 0)_{C_L[a,b]} = \sqrt{(f,f)_{C_L[a,b]}} = \sqrt{\frac{1}{b-a}\int_a^b f^2(x)dx};$$

б) пространство $R_{n+1}[a, b]$ сеточных функций, определенных в точках $x_i \in [a, b]$ ($i = 0, 1, \ldots, n$), со скалярным произведением

$$(f, \varphi)_{R_{n+1}[a,\, b]} := \frac{1}{n+1}\sum_{i=0}^{n} f(x_i)\varphi(x_i), \tag{10.9}$$

метрикой

$$\rho(f, \varphi)_{R_{n+1}[a,\, b]} := \sqrt{\frac{1}{n+1}\sum_{i=0}^{n}(f(x_i)-\varphi(x_i))^2} \tag{10.10}$$

и нормой

$$\|f\|_{R_{n+1}[a,\, b]} := \rho(f, 0)_{R_{n+1}[a,\, b]} = \sqrt{(f,f)_{R_{n+1}[a,\, b]}} = \sqrt{\frac{1}{n+1}\sum_{i=0}^{n} f^2(x_i)}.$$

Введенные указанным способом метрики (10.8) и (10.10) характеризуют близость функций $f(x)$ и $\varphi(x)$ в пространствах $C_L[a, b]$ и $R_{n+1}[a, b]$ соответственно; по отношению к приближенному равенству $f(x) \approx \varphi(x)$ при $x \in [a,b]$ они представляют собой ***интегральную*** и ***точечную*** (дискретную) ***среднеквадратические ошибки***.

В силу того, что

$$\sum_{i=0}^{n}(f(x_i)-\varphi(x_i))^2 = \min \quad \Leftrightarrow \quad \rho(f, \varphi)_{R_{n+1}[a,\, b]} = \min$$

и

$$\int_a^b (f(x)-\varphi(x))^2\,dx = \min \quad \Leftrightarrow \quad \rho(f, \varphi)_{C_L[a,\, b]} = \min,$$

можно сказать, что применение метода наименьших квадратов к аппроксимации функции $f(x)$ функцией $\varphi(x)$ заданного семейства означает подбор такой функции $\varphi(x)$, которая минимизирует среднеквадратическую погрешность приближенного равенства $f(x) \approx \varphi(x)$ в интегральном или в точечном смысле; в связи с этим, она называется ***наилучшим среднеквадратическим приближением*** $f(x)$ на заданном семействе функций.

Рассмотрим, к чему сводится процесс построения наилучших среднеквадратических приближений в одном конкретном, но достаточно общем случае, когда аппроксимирующая $f(x)$ функция $\varphi(x)$ представляет собой линейную комбинацию нескольких других, вообще говоря, более простых (базисных) функций.

Пусть $\{\varphi_j(x)\}_{j=0}^{m}$ — некоторая заданная на $[a, b]$ система линейно независимых функций. ***Обобщенным многочленом*** будем называть функцию

$$Q_m(x) := c_0\varphi_0(x) + c_1\varphi_1(x) + \ldots + c_m\varphi_m(x), \tag{10.11}$$

где $c_0, c_1, \ldots, c_m$ — произвольные вещественные числа (коэффициенты обобщенного многочлена). Поскольку функции $\varphi_j(x)$ считаются заданными, построение ***обобщенного многочлена наилучшего среднеквадратического приближения*** для данной функции $f(x)$ сводится к нахождению оптимального набора $c_0^*, c_1^*, \ldots, c_m^*$ коэффициентов $Q_m(x)$ в (10.11) на основе метода

наименьших квадратов, т.е. к решению задач минимизации:

$$\sum_{i=0}^{n}(c_0\varphi_0(x_i)+c_1\varphi_1(x_i)+\ldots+c_m\varphi_m(x_i)-f(x_i))^2\to\min \quad (10.12)$$

— в дискретном случае;

$$\int_a^b(c_0\varphi_0(x)+c_1\varphi_1(x)+\ldots+c_m\varphi_m(x)-f(x))^2\,dx\to\min \quad (10.13)$$

— в непрерывном случае.

Для первой из этих задач, т.е. для (10.12), необходимые (и достаточные) условия выражаются системой

$$\begin{cases}\sum_{i=0}^{n}\varphi_0(x_i)(c_0\varphi_0(x_i)+c_1\varphi_1(x_i)+\ldots+c_m\varphi_m(x_i)-f(x_i))=0,\\ \sum_{i=0}^{n}\varphi_1(x_i)(c_0\varphi_0(x_i)+c_1\varphi_1(x_i)+\ldots+c_m\varphi_m(x_i)-f(x_i))=0,\\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ \sum_{i=0}^{n}\varphi_m(x_i)(c_0\varphi_0(x_i)+c_1\varphi_1(x_i)+\ldots+c_m\varphi_m(x_i)-f(x_i))=0.\end{cases} \quad (10.14)$$

После элементарных преобразований она может быть переписана в терминах скалярных произведений (см. (10.10)):

$$\begin{cases}(\varphi_0,\varphi_0)c_0+(\varphi_0,\varphi_1)c_1+\ldots+(\varphi_0,\varphi_m)c_m=(\varphi_0,f),\\ (\varphi_1,\varphi_0)c_0+(\varphi_1,\varphi_1)c_1+\ldots+(\varphi_1,\varphi_m)c_m=(\varphi_1,f),\\ \cdots\cdots\cdots\cdots\cdots\cdots\\ (\varphi_m,\varphi_0)c_0+(\varphi_m,\varphi_1)c_1+\ldots+(\varphi_m,\varphi_m)c_m=(\varphi_m,f).\end{cases} \quad (10.15)$$

Если сеточные функции $\varphi_j(x_i)$ образуют систему линейно независимых элементов пространства $R_{n+1}[a,\ b]$, то полученная симметричная линейная алгебраическая система (10.15), называемая ***нормальной системой МНК***, имеет заведомо отличный от нуля определитель (это известный определитель Грама [3, 13, 57]), и значит, однозначно разрешима. Следовательно, при заданном базисе $\{\varphi_j\}$ путем решения системы (10.15) можно найти единственный обобщенный многочлен

$$Q_m^*(x):=c_0^*\varphi_0(x)+c_1^*\varphi_1(x)+\ldots+c_m^*\varphi_m(x)$$

такой, что $f(x)\approx Q_m^*(x)$ при $x\in[a,\ b]$ с наименьшей среднеквадратической ошибкой

$$\rho\left(f,\ Q_m^*(x)\right)=\sqrt{\frac{1}{n+1}\sum_{i=0}^{n}\left(f(x_i)-Q_m^*(x_i)\right)^2}\,. \quad (10.16)$$

Вернемся вновь на начальный этап построения обобщенного многочлена наилучшего среднеквадратического приближения, но теперь уже в интегральном смысле. Легко убедиться, что записав для задачи (10.13) необходимые условия, т.е. получив интегральный аналог системы (10.14), от нее приходим к той же самой нормальной системе МНК (10.15) , только скалярные произведения здесь расшифровываются с помощью формулы (10.7). Для той же $f(x)$ и при том же базисе $\{\varphi_j\}$, ввиду различий в скалярных произведениях (10.7) и (10.9), решение системы (10.15) в этом случае дает, вообще говоря, уже другой (возможно, «близкий» к $c_0^*,\ c_1^*,\ \ldots,\ c_m^*$) оптимальный набор $\tilde c_0,\ \tilde c_1,\ \ldots,\ \tilde c_m$ коэффициентов (10.11) и приводит к представлению

$$f(x)\approx\tilde Q_m(x)=\tilde c_0\varphi_0(x)+\tilde c_1\varphi_1(x)+\ldots+\tilde c_m\varphi_m(x)$$

с наименьшим среднеквадратическим отклонением

$$\rho\left(f,\ \tilde Q_m(x)\right)=\sqrt{\frac{1}{b-a}\int_a^b\left(f(x)-\tilde Q_m(x)\right)^2dx} \quad (10.17)$$

(т.е. гарантируется, что $\rho\left(f,\tilde Q_m(x)\right)_{C_L[a,b]}\le\rho\left(f,Q_m(x)\right)_{C_L[a,b]}$, где $Q_m(x)$ — произвольная функция вида (10.11)).

Замечание 10.3. Множители $\frac{1}{b-a}$ в скалярном произведении (10.7) и $\frac{1}{n+1}$ в (10.9) введены только ради удобства трактовки величин $\rho(f,\ \varphi)$ в (10.8) и в (10.10) как среднеквадратических ошибок приближенного равенства $f(x)\approx\varphi(x)$. Пользуясь более привычными формулами скалярных произведений (без этих множителей), в итоге все равно пришли бы к той же системе (10.15) (проверьте!). Так что в дальнейшем под скалярными произведениями в нормальных системах МНК могут пониматься обычные интегральные и точечные (евклидовы) скалярные произведения, в то время как при подсчете среднеквадратических погрешностей (т.е. в формулах (10.16), (10.17)) эти множители обязаны присутствовать.

Анализируя СЛАУ (10.15), приходим к выводу, что она чрезвычайно упрощается в случае, когда базисные функции $\varphi_j(x)$ образуют на $[a,\ b]$ о р т о г о н а л ь н у ю систему.

Как известно, взаимная ортогональность функций из множества $\{\varphi_j(x)\}_{j=0}^m$ означает, что $(\varphi_j,\ \varphi_k)=0$ при любых $k\ne j$. Следовательно, коэффициенты $c_0,\ c_1,\ \ldots,\ c_m$ обобщенного многочлена наилучшего среднеквадратического приближения (10.11) могут быть сразу выписаны из превратившейся в диагональную

системы (10.15), а именно:

$$\begin{cases} c_0 = \dfrac{(\varphi_0, f)}{(\varphi_0,\varphi_0)} = \dfrac{(\varphi_0, f)}{\|\varphi_0\|^2}, \\ c_1 = \dfrac{(\varphi_1, f)}{(\varphi_1,\varphi_1)} = \dfrac{(\varphi_1, f)}{\|\varphi_1\|^2}, \\ \dots\dots\dots\dots \\ c_m = \dfrac{(\varphi_m, f)}{(\varphi_m,\varphi_m)} = \dfrac{(\varphi_m, f)}{\|\varphi_m\|^2}. \end{cases} \qquad (10.18)$$

В таком случае эти оптимальные коэффициенты называются ***коэффициентами Фурье***, а обобщенный многочлен (10.11) с этим набором коэффициентов (т.е. обобщенный многочлен наилучшего среднеквадратического приближения для $f(x)$) называется ***обобщенным многочленом Фурье***.

Еще проще построение таких многочленов, когда система $\{\varphi_j(x)\}_{j=0}^m$ — ортонормированная. Тогда $\|\varphi_j\|=1$, и из (10.18) следует, что $c_j = (\varphi_j, f)$ при любом $j \in \{0, 1, \dots, m\}$, т.е. аппроксимация функции $f(x)$ обобщенным многочленом Фурье имеет вид

$$f(x) \approx (\varphi_0, f)\varphi_0(x) + (\varphi_1, f)\varphi_1(x) + \dots + (\varphi_m, f)\varphi_m(x). \qquad (10.19)$$

В частности, в математическом анализе достаточно подробно изучаются представления функций (не обязательно непрерывных) выражениями типа (10.19), в которых в качестве базисных функций используются функции ортогональной на $[-\pi, \pi]$ системы $\{\sin kx, \cos kx \mid k \in \mathbf{N}_0\}$. Такие представления в этом случае называются ***отрезками тригонометрических рядов Фурье***.

10.3. О НОРМАЛЬНОЙ СИСТЕМЕ МНК ПРИ ПОЛИНОМИАЛЬНОЙ АППРОКСИМАЦИИ

Возьмем в качестве базисных функций для обобщенного многочлена (10.11) степенные функции

$$\varphi_0 = 1, \quad \varphi_1 = x, \quad \varphi_2 = x^2, \quad \dots \quad , \quad \varphi_m = x^m.$$

В таком случае он превращается в обычный многочлен степени m канонического вида:

$$Q_m(x) \equiv P_m(x) := c_0 + c_1 x + c_2 x^2 + \dots + c_m x^m. \qquad (10.20)$$

Посмотрим, что представляет собой система (10.15) для вычисления коэффициентов многочлена $P_m(x)$, если ставится задача аппроксимировать с его помощью некоторую функцию $f(x)$ по метрике пространства $C_L[0, 1]$, т.е. для нахождения такого набора коэффициентов $c_0, c_1, \dots, c_m$ в (10.20), при котором обеспечивается минимум величины

$$\int_0^1 (P_m(x) - f(x))^2\, dx.$$

Подсчитав скалярное произведение вида (10.7)

$$(\varphi_i, \varphi_j) = \int_0^1 x^i \cdot x^j dx = \int_0^1 x^{i+j} dx = \frac{1}{i+j+1},$$

из (10.15), варьируя i и j от 0 до m, получаем систему

$$\begin{cases} c_0 + \dfrac{1}{2}c_1 + \dots + \dfrac{1}{m+1}c_m = \int_0^1 f(x)dx, \\ \dfrac{1}{2}c_0 + \dfrac{1}{3}c_1 + \dots + \dfrac{1}{m+2}c_m = \int_0^1 x f(x)dx, \\ \dots\dots\dots\dots\dots\dots \\ \dfrac{1}{m+1}c_0 + \dfrac{1}{m+2}c_1 + \dots + \dfrac{1}{2m+1}c_m = \int_0^1 x^m f(x)dx. \end{cases} \qquad (10.21)$$

Эта нормальная система МНК для построения многочлена m-й степени наилучшего среднеквадратического приближения функции $f(x)$ на отрезке $[0, 1]$ в векторно-матричных обозначениях имеет вид

$$\mathbf{H}_{m+1} \cdot \mathbf{c} = \mathbf{r},$$

где $\mathbf{c} := (c_0; c_1; \dots; c_m)^T$ — вектор неизвестных (коэффициентов многочлена $P_m(x)$ в (10.20)),

$$\mathbf{r} := \begin{pmatrix} \int_0^1 f(x)dx \\ \int_0^1 x f(x)dx \\ \dots\dots \\ \int_0^1 x^m f(x)dx \end{pmatrix} \quad \text{и} \quad \mathbf{H}_{m+1} := \left(\frac{1}{i+j-1}\right)_{i,\,j=1}^{m+1}$$

— вектор свободных членов и матрица коэффициентов СЛАУ (10.21) соответственно.

Матрица $\mathbf{H}_{m+1}$ носит название ***матрицы Гильберта*** и ши-

14 Б-223

роко известна в математике как классический пример плохообусловленной матрицы. Элементы этой матрицы являются простыми функциями индексов, она заведомо не вырождена и, в то же время, ее ***мера обусловленности*** $\operatorname{cond}\mathbf{H}_{m+1}$ с увеличением m растет чрезвычайно быстро [3, 183 и др.]. Так, если при $m=1$ можно считать $\operatorname{cond}\mathbf{H}_2 = \|\mathbf{H}_2\| \cdot \left\|\mathbf{H}_2^{-1}\right\| \approx 20$, то уже при $m=5$ имеем $\operatorname{cond}\mathbf{H}_6 \approx 10^7$, а при $m=9$ $\operatorname{cond}\mathbf{H}_{10} \approx 10^{13}$.

Зная, что число обусловленности характеризует чувствительность СЛАУ к ошибкам в исходных данных, т.е. фактически служит коэффициентом увеличения относительных погрешностей, которые неизбежны при переводе обыкновенных дробей в десятичные и при вычислении интегралов в правой части системы (10.21) в процессе ее решения, можно сделать вывод о непригодности использования базиса из степенных функций для построения многочленов наилучших среднеквадратических приближений сколько-нибудь высоких степеней.

Если рассматривать дискретный случай, т.е. строить с помощью МНК аппроксимирующий $f(x)$ при $x \in [0, 1]$ многочлен $P_m(x)$ (10.20), исходя из знания $n+1$ значений $f(x_i)$ в точках $x_0, x_1, \ldots, x_n$ отрезка $[0, 1]$, то подстановка в систему (10.15) полученных на основе определения скалярного произведения (10.9) выражений

$$(\varphi_k, \varphi_j) = \frac{1}{n+1}\sum_{i=0}^{n} x_i^{k+j}, \quad (\varphi_k, f) = \frac{1}{n+1}\sum_{i=0}^{n} x_i^k f(x_i)$$

приводит к нормальной системе МНК вида

$$\begin{cases} c_0 + \left(\frac{1}{n+1}\sum_{i=0}^{n} x_i\right)c_1 + \ldots + \left(\frac{1}{n+1}\sum_{i=0}^{n} x_i^m\right)c_m = \frac{1}{n+1}\sum_{i=0}^{n} f(x_i), \\ \left(\frac{1}{n+1}\sum_{i=0}^{n} x_i\right)c_0 + \left(\frac{1}{n+1}\sum_{i=0}^{n} x_i^2\right)c_1 + \ldots + \left(\frac{1}{n+1}\sum_{i=0}^{n} x_i^{m+1}\right)c_m = \\ \qquad = \frac{1}{n+1}\sum_{i=0}^{n} x_i f(x_i), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \left(\frac{1}{n+1}\sum_{i=0}^{n} x_i^m\right)c_0 + \left(\frac{1}{n+1}\sum_{i=0}^{n} x_i^{m+1}\right)c_1 + \ldots + \left(\frac{1}{n+1}\sum_{i=0}^{n} x_i^{2m+1}\right)c_m = \\ \qquad = \frac{1}{n+1}\sum_{i=0}^{n} x_i^m f(x_i). \end{cases} \tag{10.22}$$

Разумеется, множитель $\frac{1}{n+1}$ в каждом уравнении можно опустить, однако с ним легче увидеть, что, например, в случае равномерного распределения узлов x_i на отрезке $[0, 1]$, т.е. при $x_i = \frac{i}{n}$, матрица коэффициентов полученной системы (10.22) будет асимптотически приближаться с ростом n к матрице системы (10.21) (т. е. к матрице Гильберта $\mathbf{H}_{m+1}$), что порождает описанные выше проблемы при решении системы, связанные с численной неустойчивостью.

Заметим, что если линейная независимость системы степенных функций $1, x, \ldots, x^m$ как элементов пространства $C_L[0, 1]$ хорошо известна и фактически видна из невырожденности матрицы Гильберта системы (10.21), то с линейной независимостью системы их сеточных аналогов стоит разобраться.

На сетке несовпадающих узлов $x_i \in [0, 1]$ при $i = 0, 1, \ldots, n$ из непрерывных степенных функций $1, x, \ldots, x^m$ получаем систему из $m+1$ векторов размерности $n+1$ — элементов пространства $R_{n+1}[0, 1]$ сеточных функций:

$$\varphi_0 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \varphi_1 = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \varphi_2 = \begin{pmatrix} x_0^2 \\ x_1^2 \\ \vdots \\ x_n^2 \end{pmatrix}, \quad \ldots, \quad \varphi_m = \begin{pmatrix} x_0^m \\ x_1^m \\ \vdots \\ x_n^m \end{pmatrix}.$$

Эти элементы образуют линейно независимую систему тогда и только тогда, когда равенство нулю их линейной комбинации $\alpha_0\varphi_0 + \alpha_1\varphi_1 + \alpha_2\varphi_2 + \ldots + \alpha_m\varphi_m$ возможно только при нулевых значениях коэффициентов $\alpha_0, \alpha_1, \ldots, \alpha_m$, т.е. если однородная линейная относительно $\alpha_0, \alpha_1, \ldots, \alpha_m$ СЛАУ

$$\begin{cases} \alpha_0 + x_0\alpha_1 + x_0^2\alpha_2 + \ldots + x_0^m\alpha_m = 0, \\ \alpha_0 + x_1\alpha_1 + x_1^2\alpha_2 + \ldots + x_1^m\alpha_m = 0, \\ \ldots\ldots\ldots\ldots\ldots \\ \alpha_0 + x_n\alpha_1 + x_n^2\alpha_2 + \ldots + x_n^m\alpha_m = 0 \end{cases}$$

имеет только тривиальное решение. Последнее возможно только при условии, что ранг матрицы системы равен $m+1$, но для этого нужно, чтобы число уравнений было не меньше, чем $m+1$. Отсюда приходим к необходимости требования, чтобы *число то-*

чек согласования $(n+1)$ *было не меньшим числа отыскиваемых параметров* $(m+1)$. При этом очевидно, что случай $m=n$ замыкает теорию полиномиальной аппроксимации по методу наименьших квадратов на теорию лагранжевой интерполяции.

10.4. СИСТЕМЫ ОРТОГОНАЛЬНЫХ МНОГОЧЛЕНОВ

Имея в виду, что нас интересует, в первую очередь, полиномиальная аппроксимация функций, и при этом показаны недостатки степенных функций для их использования в качестве базисных при построении многочленов наилучших среднеквадратических приближений (§ 10.3) и отмечены достоинства ортогональных многочленов в этой роли (§ 10.2), приведем некоторые сведения справочного характера о последних.

Многочлены Лежандра $\chi_n(x)$ — наиболее употребительные из классических ортогональных многочленов и единственные, для которых условие их ортогональности на отрезке $[-1, 1]$ выполняется «в чистом виде», т.е. через равенство нулю скалярного произведения вида

$$(f, \varphi) := \int_{-1}^{1} f(x)\varphi(x)dx, \qquad (10.23)$$

а именно:

$$\int_{-1}^{1} \chi_k(x)\chi_j(x)dx = \begin{cases} 0, & \text{если } k \neq j, \\ \dfrac{2}{2k+1}, & \text{если } k = j. \end{cases} \qquad (10.24)$$

Для многочленов Лежандра известна явная, порождающая их, ***формула Родрига****)

$$\chi_n(x) := \frac{1}{n!\,2^n} \cdot \frac{d^n}{dx^n}\left(x^2-1\right)^n. \qquad (10.25)$$

Положив $\chi_0 = 1$ и найдя по формуле (10.25) $\chi_1 = x$, все последующие многочлены Лежандра можно получать один за другим с помощью рекуррентной формулы

$$(n+1)\chi_{n+1}(x) - (2n+1)x\,\chi_n(x) + n\chi_{n-1}(x) = 0. \qquad (10.26)$$

*) Родриг (Родригес) Бенжамен Оленд (1794–1851) — французский математик. Появление формулы Родрига (10.25) датируется 1814 годом.

При $n = 1, 2, 3, \ldots$ из нее имеем соответственно:

$$\chi_2 = \frac{1}{2}\left(3x^2 - 1\right),$$

$$\chi_3 = \frac{1}{2}\left(5x^3 - 3x\right),$$

$$\chi_4 = \frac{1}{8}\left(35x^4 - 30x^2 + 3\right),$$

$$\chi_5 = \frac{1}{8}\left(63x^5 - 70x^3 + 15x\right) \qquad \text{и т. д.}$$

Сравнивая многочлены Лежандра с рассмотренными в гл. 9 многочленами Чебышева, можно обнаружить, что их сходство — не только внешнее. Эти многочлены характеризует ряд одинаковых свойств. Например, как и многочлены Чебышева, многочлены Лежандра n-й степени имеют на отрезке $[-1, 1]$ ровно n различных действительных корней (см. свойство 9.1). В то же время, как и многочлены Лежандра, многочлены Чебышева $T_n(x)$ относятся к числу классических ортогональных многочленов, но ортогональность здесь понимается в более широком смысле — это ***ортогональность с весом***. А именно, условие ортогональности многочленов Чебышева $T_n(x)$ на отрезке $[-1, 1]$ имеет вид

$$\int_{-1}^{1} T_k(x)\cdot T_j(x)\frac{dx}{\sqrt{1-x^2}} = \begin{cases} 0, & \text{если } k \neq j, \\ \dfrac{\pi}{2}, & \text{если } k = j \neq 0, \\ \pi, & \text{если } k = j = 0. \end{cases}$$

Наряду с многочленами $T_n(x)$, называемыми также ***многочленами Чебышева первого рода***, ортогональными на $(-1, 1)$ являются и ***многочлены Чебышева второго рода*** $U_n(x)$, определяемые через $T_n(x)$, а именно:

$$U_n(x) := \frac{1}{n+1}T'_{n+1}(x) = \frac{\sin[(n+1)\arccos x]}{\sin \arccos x}.$$

Многочлены Лежандра и многочлены Чебышева первого и второго рода принадлежат одному семейству ***многочленов Якоби***, ортогональных на конечном промежутке (a, b) с некоторой весовой функцией $p(x) > 0$ определенного вида (для многочленов Лежандра $p(x) \equiv 1$) и задаваемых формулой, обобщающей формулу Родрига (10.25) [132 и др.].

Многочлены Лагерра*) $L_n(x)$ определим требованием их ортогональности на промежутке $[0, +\infty)$ с весовой функцией e^{-x}, а именно, условием

$$\int_0^{+\infty} e^{-x} L_k(x) \cdot L_j(x) dx = \begin{cases} 0 & ,\text{если } k \neq j, \\ (k!)^2 & ,\text{если } k = j. \end{cases}$$

Как и предыдущие ортогональные многочлены, многочлены Лагерра удовлетворяют рекуррентному соотношению типа (10.26)

$$L_{n+1}(x) - (2n+1-x)L_n(x) + n^2 L_{n-1}(x) = 0, \qquad (10.27)$$

где $n \in \mathbf{N}$, $L_0 := 1$, $L_1 := -x+1$. Отсюда легко получить несколько первых многочленов Лагерра:

$L_2 = x^2 - 4x + 2$,
$L_3 = -x^3 + 9x^2 - 18x + 6$,
$L_4 = x^4 - 16x^3 + 72x^2 - 96x + 24$,
$L_5 = -x^5 + 25x^4 - 200x^3 + 600x^2 - 600x + 120$ и т. д.

Заметим, что под тем же обозначением $L_n(x)$ в литературе может встретиться многочлен $\frac{1}{n!} L_n(x)$. Кроме того, следует упомянуть, что существуют ортогональные многочлены Лагерра несколько более общего вида; в частности, ортогональность может пониматься с весовой функцией $p(x) = x^{\alpha} e^{-x}$, где $\alpha > -1$ (в рассмотренном случае $\alpha = 0$).

Многочлены Эрмита**) $H_n(x)$ ортогональны на всей числовой оси с весом $p(x) = e^{-x^2}$. Они удовлетворяют интегральному условию ортогональности

$$\int_{-\infty}^{+\infty} e^{-x^2} H_k(x) \cdot H_j(x) dx = \begin{cases} 0, & \text{если } k \neq j, \\ 2^k k! \sqrt{\pi}, & \text{если } k = j, \end{cases}$$

и рекуррентному соотношению

$$H_{n+1}(x) - 2xH_n(x) + 2nH_{n-1}(x) = 0, \qquad (10.28)$$

*) Лагéрр Эдмон Никола (1834–1886) — французский математик.

**) Не надо путать определяемые здесь ортогональные многочлены Эрмита с рассмотренными ранее интерполяционными многочленами Эрмита (см. § 8.8). Обозначение $H_n(x)$ (как и $L_n(x)$) несет в этой книге двойную нагрузку.

где $n \in \mathbf{N}$, $H_0 := 1$, $H_1 := 2x$. Подставляя в (10.28) последовательно $n = 1, 2, \ldots$, находим первые многочлены Эрмита:

$H_2 = 4x^2 - 2$,
$H_3 = 8x^3 - 12x$,
$H_4 = 16x^4 - 48x^2 + 12$,
$H_5 = 32x^5 - 160x^3 + 120x$ и т. д.

Как видим, они имеют такую же структуру, как и многочлены Чебышева.

Ортогональные многочлены Якоби, Лагерра и Эрмита обладают рядом общих свойств. Все они являются решениями одного семейства обыкновенных дифференциальных уравнений второго порядка. Все ортогональные многочлены $P_n(x)$ одной и той же системы линейно независимы и удовлетворяют трехчленному рекуррентному соотношению вида

$$\alpha_n P_{n+1}(x) + (\beta_n - x)P_n(x) + \gamma_n P_{n-1}(x) = 0, \qquad (10.29)$$

содержащему в себе, как легко убедиться, соотношения (10.26), (10.27), (10.28), а также (9.2), для конкретных многочленов Лежандра, Лагерра, Эрмита и Чебышева соответственно. Все n корней ортогональных многочленов $P_n(x)$ — простые и находятся на промежутке ортогональности; при этом корни многочлена $P_n(x)$ разделяются корнями многочлена $P_{n-1}(x)$ [187].

Более подробные сведения об ортогональных многочленах и их свойствах можно найти в специальной, справочной, учебной и даже научно-популярной литературе [19, 90, 109, 132, 169, 187].

10.5. ПРОСТАЯ ПРОЦЕДУРА ПОСТРОЕНИЯ СИСТЕМЫ ОРТОГОНАЛЬНЫХ МНОГОЧЛЕНОВ

В целях полиномиальной аппроксимации функций может оказаться более удобным использование не конкретных классических ортогональных многочленов, введенных в предыдущем параграфе и жестко привязанных к конкретному скалярному произведению (они еще потребуются в дальнейшем, см. гл. 12), а последовательное построение системы многочленов, взаимно ортогональных в смысле того или иного скалярного произведения $(\bullet, \bullet)$, фиксируемого в конкретных ситуациях.

Итак, будем строить такую систему многочленов $\{q_k\}_{k=0}^m$, что:

1) $q_0 := 1$;

2) $q_k = q_k(x)$ — многочлен степени k с коэффициентом 1 при старшей степени x;

3) $(q_k, q_j) = 0$ при любых $j \neq k$ (при этом заведомо $(q_k, q_k) = \|q_k\|^2 \neq 0$).

Эта система будет получаться последовательным подсоединением к заданному ее элементу $q_0 := 1$ многочленов повышающихся на единицу степеней, ортогональных всем предыдущим.

Полагая $q_1 := x - \alpha_1$, из условия $q_1 \perp q_0$, т.е. из $(q_1, q_0) = 0$, по свойствам скалярного произведения имеем

$$(x, q_0) - \alpha_1 (q_0, q_0) = 0,$$

откуда получаем выражение коэффициента α_1:

$$\alpha_1 = \frac{(x, q_0)}{(q_0, q_0)}. \tag{10.30}$$

Далее воспользуемся тем фактом, что рассмотренные выше классические ортогональные многочлены удовлетворяют трехчленному рекуррентному соотношению (10.29) с некоторыми наборами параметров α_n, β_n, γ_n. В силу договоренности о том, что здесь будут строиться ортогональные многочлены $q_k(x)$ со старшим членом x^k, можно ограничиться меньшим числом параметров и попытаться находить $(k+1)$-й многочлен $q_{k+1}(x)$ по уже известным многочленам $q_{k-1}(x)$ и $q_k(x)$ по формуле

$$q_{k+1} = x q_k - \alpha_{k+1} q_k - \beta_k q_{k-1}, \tag{10.31}$$

которая будет полностью определять дальнейший процесс построения $q_2, q_3, q_4, \ldots$, если будут известны законы вычисления коэффициентов α_{k+1} и β_k при $k = 1, 2, 3, \ldots$.

Сначала убедимся, что формула (10.31) «работает» при $k = 1$. Для этого нужно показать, что в равенстве $q_2 = x q_1 - \alpha_2 q_1 - \beta_1 q_0$ можно однозначно найти коэффициенты α_2 и β_1 так, чтобы многочлен q_2 был ортогонален одновременно многочленам q_0 и q_1. Имеем:

$$\begin{cases} (q_2, q_0) = 0, \\ (q_2, q_1) = 0 \end{cases} \Leftrightarrow \begin{cases} (x q_1, q_0) - \alpha_2 (q_1, q_0) - \beta_1 (q_0, q_0) = 0, \\ (x q_1, q_1) - \alpha_2 (q_1, q_1) - \beta_1 (q_0, q_1) = 0. \end{cases}$$

Но многочлен q_1, благодаря (10.30), таков, что $(q_1, q_0) = 0$; поэтому из последней системы сразу получаем

$$\beta_1 = \frac{(x q_1, q_0)}{(q_0, q_0)}, \quad \alpha_2 = \frac{(x q_1, q_1)}{(q_1, q_1)}.$$

Предположим, что уже построены многочлены $q_0, q_1, \ldots, q_k$ такие, что имеет место их попарная ортогональность:

$$q_i \perp q_j \quad \forall\, i, j \in \{0, 1, \ldots, k\} : i \neq j. \tag{10.32}$$

Беря многочлен q_{k+1} в форме (10.31), покажем, что он заведомо ортогонален каждому из многочленов от q_0 до q_{k-2} независимо от значений коэффициентов α_{k+1} и β_k, и найдем эти значения из требования ортогональности многочлена q_{k+1} многочленам q_{k-1} и q_k.

Для этого сначала рассматриваем скалярные произведения (q_{k+1}, q_j) при $j = 0, 1, \ldots, k-2$:

$$(q_{k+1}, q_j) = (x q_k, q_j) - \alpha_{k+1} (q_k, q_j) - \beta_k (q_{k-1}, q_j).$$

В этом равенстве имеем $(q_k, q_j) = 0$ и $(q_{k-1}, q_j) = 0$, в силу договоренности (10.32), а равенство нулю скалярного произведения $(x q_k, q_j)$ можно установить на основе равенства $(x q_k, q_j) = (q_k, x q_j)$, очевидного, по крайней мере, для используемых здесь скалярных произведений вида (10.7), (10.9) (для которых в выражении скалярного произведения степенных функций (x^k, x^j) применимо их свойство $x^k \cdot x^j = x^{k+j}$). Действительно, так как при $j \leq k-2$ степень многочлена $x q_j$ не превосходит $k-1$ и так как ортогональные многочлены q_j линейно независимы, то многочлен $x q_j$ может быть представлен в виде линейной комбинации многочленов $q_0, q_1, \ldots, q_{j+1}$ с некоторыми коэффициентами $\gamma_0, \gamma_1, \ldots, \gamma_{j+1}$, и значит,

$$(q_k, x q_j) = \gamma_0 (q_k, q_0) + \gamma_1 (q_k, q_1) + \ldots + \gamma_{j+1} (q_k, q_{j+1}) = 0,$$

поскольку $j + 1 \leq k - 1$ и выполняется условие (10.32).

Далее, из требования $(q_{k+1}, q_{k-1}) = 0$, т.е. из

$$(x q_k, q_{k-1}) - \alpha_{k+1} (q_k, q_{k-1}) - \beta_k (q_{k-1}, q_{k-1}) = 0,$$

учитывая, что $(q_k, q_{k-1}) = 0$, находим

$$\beta_k = \frac{(x q_k, q_{k-1})}{(q_{k-1}, q_{k-1})}, \tag{10.33}$$

и из требования $(q_{k+1}, q_k) = 0$, т.е. из

$$(x q_k, q_k) - \alpha_{k+1} (q_k, q_k) - \beta_k (q_{k-1}, q_k) = 0,$$

аналогично получаем

$$\alpha_{k+1} = \frac{(xq_k, q_k)}{(q_k, q_k)}. \qquad (10.34)$$

Таким образом, завязывая воедино выведенные для выражения (10.31) формулы коэффициентов (10.30), (10.33) и (10.34), приходим к следующей *итерационной процедуре получения ортогональных многочленов все повышающихся степеней*:

$$q_{k+1}(x) = xq_k(x) - \frac{(xq_k(x), q_k(x))}{(q_k(x), q_k(x))} q_k(x) - \frac{(xq_k(x), q_{k-1}(x))}{(q_{k-1}(x), q_{k-1}(x))} q_{k-1}(x), \qquad (10.35)$$

где $k = 1, 2, \ldots;\quad q_0(x) \equiv 1,\quad q_1(x) = x - \dfrac{(x, q_0(x))}{(q_0(x), q_0(x))}$.

10.6. АППРОКСИМАЦИЯ ФУНКЦИЙ МНОГОЧЛЕНАМИ ФУРЬЕ

Резюмируя результаты §§ 10.2 и 10.5, приходим к выводу, что построение многочлена m-й степени $Q_m(x)$, осуществляющего ***наилучшее среднеквадратическое приближение*** для заданной функции $f(x)$, можно выполнить достаточно просто и численно устойчиво, представляя ее отрезком ряда Фурье:

$$f(x) \approx Q_m(x) = \sum_{j=0}^{m} c_j q_j(x), \qquad (10.36)$$

где попарно ортогональные многочлены $q_j(x)$ можно получать итерационным процессом (10.35), а коэффициенты Фурье суть (см.(10.18))

$$c_j = \frac{(q_j(x), f(x))}{(q_j(x), q_j(x))}. \qquad (10.37)$$

Те или иные конкретные аппроксимации $f(x)$ типа (10.36) получаются фиксированием тех или иных конкретных скалярных произведений.

Например, аппроксимируя функцию $f(x)$ на отрезке $[a, b]$ многочленом $Q_m(x)$ (10.36) таким образом, чтобы минимизировалось определенное равенством (10.8) среднеквадратическое

отклонение многочлена $Q_m(x)$ от данной функции $f(x)$, в формулах (10.35) и (10.37) нужно использовать интегральное скалярное произведение (10.7) (с множителем $\frac{1}{b-a}$ или без него, что не играет роли). В результате в представлении (10.36) многочлены $q_j(x)$ должны получаться последовательно, начиная с $q_0(x) \equiv 1$,

$$q_1(x) = x - \frac{\int_a^b x\,dx}{\int_a^b dx} = x - \frac{a+b}{2}, \quad \text{рекуррентным равенством}$$

$$q_{k+1}(x) = xq_k(x) - \frac{\int_a^b xq_k^2(x)dx}{\int_a^b q_k^2(x)dx} q_k(x) - \frac{\int_a^b xq_k(x)q_{k-1}(x)dx}{\int_a^b q_{k-1}^2(x)dx} q_{k-1}(x)$$

при $k = 1, 2, \ldots, m-1$, а коэффициенты c_j должны вычисляться по формуле

$$c_j = \frac{\int_a^b q_j(x)f(x)dx}{\int_a^b q_j^2(x)dx} \quad \text{при} \quad j = 0, 1, \ldots, m.$$

Рассмотрим теперь случай, когда функция $f(x)$ задана т а б л и ц е й своих значений в точках $x_0 < x_1 < \ldots < x_n$ отрезка $[a, b]$. Для аппроксимации этой сеточной функции многочленом Фурье степени m при $m \le n$ на основе формул (10.35)–(10.37) используем евклидово скалярное произведение (10.9) (опять-таки с множителем $\frac{1}{n+1}$ или без него), согласно которому, например,

$$(q_0, q_0) = \sum_{i=0}^{n} 1 = n+1, \quad (x, q_0) = \sum_{i=0}^{n} x_i, \text{ т.е. } \alpha_1 = \frac{1}{n+1}\sum_{i=0}^{n} x_i, \text{ и т.д.}$$

Если при этом степень m аппроксимирующего многочлена $Q_m(x)$ жестко зафиксирована, то этот случай не имеет принципиального отличия от предыдущего. А именно, начиная с

$$q_0(x) \equiv 1, \quad q_1(x) = x - \frac{1}{n+1}\sum_{i=0}^{n} x_i, \quad c_0 = \frac{1}{n+1}\sum_{i=0}^{n} f(x_i), \qquad (10.38)$$

при $k = 1, 2, \ldots, m-1$ вычисляем коэффициенты

$$\alpha_{k+1} = \frac{\sum_{i=0}^{n} x_i q_k^2(x_i)}{\sum_{i=0}^{n} q_k^2(x_i)}, \quad \beta_k = \frac{\sum_{i=0}^{n} x_i q_k(x_i) q_{k-1}(x_i)}{\sum_{i=0}^{n} q_{k-1}^2(x_i)} \qquad (10.39)$$

и с их помощью конструируем многочлены $q_2(x)$, $q_3(x)$, ..., $q_m(x)$ по формуле

$$q_{k+1}(x) = xq_k(x) - \alpha_{k+1}q_k(x) - \beta_k q_{k-1}(x); \qquad (10.40)$$

затем при $j = 1, 2, \ldots, m$ вычисляем коэффициенты Фурье

$$c_j = \frac{\sum_{i=0}^{n} q_j(x_i) f(x_i)}{\sum_{i=0}^{n} q_j^2(x_i)} \qquad (10.41)$$

для их подстановки в выражение (10.36).

Заметим, что при $m = n$ получаемый таким образом многочлен Фурье $Q_n(x)$ наилучшего среднеквадратического приближения сеточной функции $f(x)$ на $[a, b]$ должен совпасть с соответствующим интерполяционным многочленом Лагранжа $L_n(x)$ (см. (1.6)). Следовательно, *многочлен*

$$Q_n(x) = \sum_{j=0}^{n} c_j q_j(x)$$

с *определенными с помощью формул* (10.38)–(10.41) *коэффициентами* c_j *и ортогональными многочленами* $q_j(x)$ *дает еще одну форму представления интерполяционного многочлена наряду с несколькими другими, изученными в гл.* 8.

Пусть теперь речь идет не о представлении функции $f(x)$, заданной таблицей своих приближенных значений $f(x_i)$ ($i = 0, 1, \ldots, n$), многочленом $Q_m(x)$ определенной степени m, а о вычислении ее приближенного значения $f(\tilde{x})$ в заданной точке $\tilde{x} \in [a, b]$, причем считается, что $a \approx x_0 < x_1 < \ldots < x_n \approx b$ и $\tilde{x} \neq x_i$.

Учитывая, что повышение степени m многочлена Фурье $Q_m(x)$, аппроксимирующего $f(x)$ на $[a, b]$, улучшает качество аппроксимации, т. е. уменьшает величину среднеквадратической ошибки (10.16) до тех пор, пока на погрешность метода не станет существенно влиять погрешность округлений (вычислений), *есть смысл совместить процесс вычисления коэффициентов Фурье с процедурой ортогонализации* в одном цикле и производить последовательное вычисление значений $Q_0(\tilde{x})$, $Q_1(\tilde{x})$, ..., все более точно приближающих значение $f(\tilde{x})$. Алгоритм, реализующий такой подход, должен предусматривать вычисление на каждом k-м шаге среднеквадратических отклонений

$$\rho_k = \sqrt{\frac{1}{n+1}\sum_{i=0}^{n}(f(x_i) - Q_k(x_i))^2}$$

и проверку их на убывание при переходе от k к $k+1$, проверку величин $Q_k(\tilde{x})$ на сближение при увеличении k (т.е. того, что $|Q_{k+1}(\tilde{x}) - Q_k(\tilde{x})| < |Q_k(\tilde{x}) - Q_{k-1}(\tilde{x})|$), а также того, что $k < n$.

Кстати, при небольшом числе узлов, т.е. при небольших значениях n, случай $k = n$ можно использовать для проверки правильности работы алгоритма. Действительно, в этом случае многочлен $Q_k(x) \equiv Q_n(x)$ должен стать интерполяционным, но поскольку он строился из других соображений, сравнение величины $f(x_i) - Q_n(x_i)$ или ρ_n с нулем не тривиально (не то, что при классической интерполяции).

УПРАЖНЕНИЯ

10.1. Найдите нормальное псевдорешение системы

$$\begin{cases} 2x_1 + 3x_2 - 4x_3 = 1.2, \\ 3x_1 + x_2 + 7x_3 = 4.7, \\ x_1 - 4x_2 - 3x_3 = 6.1, \\ 4x_1 + 3x_2 - 5x_3 = 5.2, \\ 2x_1 - 4x_2 + 3x_3 = 7.9. \end{cases}$$

10.2. Для функции $y = f(x)$, заданной таблицей

x	0.5	1	1.5	2	2.5
y	10.5	1.6	0.55	0.26	0.15

подберите подходящий вид аппроксимирующей ее нелинейной зависимости из следующих:

а) $y = a \cdot x^b$; б) $y = a \cdot e^{bx}$; в) $y = a + \frac{b}{x}$; г) $y = \frac{1}{a + bx}$,

находя методом наименьших квадратов их параметры и сравнивая между собой среднеквадратические погрешности.

10.3. Постройте наилучшие среднеквадратические линейные аппроксимации для функций

а) $f(x)=\sqrt{x}$ при $x\in[0,1]$,

б) $f(x)=\dfrac{1}{x}$ при $x\in[1,2]$,

в) $f(x)=\ln(1+x)$ при $x\in[0,1]$

непосредственным применением метода наименьших квадратов. Сравните полученные результаты с результатами чебышевской аппроксимации (см. упр.9.4):

А) по величине среднеквадратической погрешности;

Б) по величине максимальной погрешности на заданном отрезке.

10.4. Функция $y=f(x)$ задана таблицей

x	0	0.1	0.2	0.3	0.4	0.5	0.6	0.8	0.9	1.0
y	0	0.095	0.182	0.262	0.337	0.406	0.470	0.588	0.642	0.693

Методом наименьших квадратов аппроксимируйте данную функцию $f(x)$ функцией $\varphi(x)=a_0+a_1x$ по трем точкам ($x=0$, $x=0.5$ и $x=1.0$), по шести ($x=0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1.0$) и по всем одиннадцати точкам. Сравните результаты (в разных метриках) с результатом применения интегрального МНК, зная, что $f(x)=\ln(1+x)$ (см. упр. 10.3, в).

10.5. Покажите, что матрица системы (10.22) при $x_i=\dfrac{i}{n}\in[0,1]$, $m=1$ и $n\to\infty$ совпадает с матрицей Гильберта $\mathbf{H}_2$.

10.6. Для функций а–в упр.10.3 на заданных отрезках постройте многочлены Фурье первой и второй степени. Найдите среднеквадратические ошибки результатов. Совпадают ли построенные многочлены первой степени с линейными функциями, полученными в упр. 10.3?

ГЛАВА 11 ‖ ИНТЕРПОЛЯЦИОННЫЕ СПЛАЙНЫ*)

Рассматривается высокотехнологичный способ приближения таблично заданных функций с помощью составных функций, звеньями которых служат многочлены небольших степеней, допускающие гладкую стыковку. Наряду со ставшими уже классическими естественными кубическими сплайнами, здесь изучаются интерполяционные параболические (квадратичные) сплайны, а также базисные сплайны (иначе, B-сплайны) и эрмитовы (локальные) сплайны нечетных и четных степеней. Дается представление о сглаживании эмпирических данных посредством простейших линейных фильтров и указывается на возможность построения сглаживающих сплайнов.

11.1. КУСОЧНО-ПОЛИНОМИАЛЬНАЯ АППРОКСИМАЦИЯ. ЛИНЕЙНЫЕ ФИЛЬТРЫ

В тех случаях, когда промежуток $[a,\ b]$, на котором нужно подменить функцию $f(x)$ функцией $\varphi(x)$, велик, и отсутствуют основания считать данную функцию $f(x)$ достаточно гладкой при $x\in[a,\ b]$, нет смысла пытаться повышать качество ее полиномиальной аппроксимации за счет использования в роли $\varphi(x)$ многочленов высоких степеней. Более перспективным в этих условиях является применение ***кусочно-полиномиальной аппроксимации*** $f(x)$, предполагающей, что аппроксимирующая функция $\varphi(x)$ составляется из отдельных многочленов, как правило, одинаковой небольшой степени, определенных каждый на своей части отрезка $[a,\ b]$. При этом, если функция $f(x)$ непрерывна и имеется достаточное количество точечной информации о ней, то можно рассчитывать приблизить ее на $[a,\ b]$ сколь угодно хорошо кусочно-полиномиальной функцией $\varphi(x)$ только за счет увеличения числа частичных промежутков, составляющих $[a,\ b]$,

---

*) Процитируем абзац, с которого начинается посвященный сплайнам параграф в книге [13]:

«Способ приближения сплайнами интересен, кроме всего прочего, отношением к нему специалистов. Одни считают его универсальным методом решения проблем, стоящих перед численным анализом, и ищут применения ему в самых различных направлениях. Другие рассматривают его как очередную дань переменчивой моде. По-видимому, истина находится где-то посередине».

при любых фиксированных степенях составных многочленов и любых способах согласования $f(x)$ и $\varphi(x)$.

Использование низких степеней многочленов, составляющих $\varphi(x)$, позволяет легко находить их коэффициенты как из интерполяционных, так и из иных условий.

Так, если заданы значения y_i функции $y = f(x)$ на системе узлов x_i таких, что

$$a \le x_0 < x_1 < \ldots < x_n \le b, \qquad (11.1)$$

и требуется аппроксимировать $f(x)$ ***кусочно-линейной функцией*** $\varphi(x)$, исходя из условий интерполяции

$$\varphi(x_i) = y_i, \quad (i = 0, 1, \ldots, n),$$

то, беря функцию $\varphi(x)$ в виде

$$\varphi(x) = \begin{cases} a_1 x + b_1 & \text{при } x \in [x_0, x_1], \\ a_2 x + b_2 & \text{при } x \in [x_1, x_2], \\ \cdots\cdots\cdots & \\ a_n x + b_n & \text{при } x \in [x_{n-1}, x_n], \end{cases} \qquad (11.2)$$

для нахождения n пар ее коэффициентов a_k, b_k $(k = 1, 2, \ldots, n)$ имеем систему из $2n$ линейных уравнений

$$\begin{cases} \begin{cases} a_1 x_0 + b_1 = y_0, \\ a_1 x_1 + b_1 = y_1; \end{cases} \\ \begin{cases} a_2 x_1 + b_2 = y_1, \\ a_2 x_2 + b_2 = y_2; \end{cases} \\ \cdots\cdots\cdots \\ \begin{cases} a_n x_{n-1} + b_n = y_{n-1}, \\ a_n x_n + b_n = y_n, \end{cases} \end{cases} \qquad (11.3)$$

причем, как видим, каждая пара соседних уравнений системы (11.3), имеющих коэффициенты с одинаковыми индексами, не связана с остальными и может решаться отдельно.

Аналогично, каждое звено ***кусочно-квадратичной функции*** (при $n = 2m$ в (11.1))

$$\varphi(x) = \begin{cases} a_1 x^2 + b_1 x + c_1 & \text{при } x \in [x_0, x_2], \\ a_2 x^2 + b_2 x + c_2 & \text{при } x \in [x_2, x_4], \\ \cdots\cdots\cdots & \\ a_m x^2 + b_m x + c_m & \text{при } x \in [x_{2m-2}, x_{2m}] \end{cases} \qquad (11.4)$$

определяется тройкой коэффициентов a_k, b_k, c_k $(k = 1, 2, \ldots, m)$, которые могут быть найдены последовательным решением (при $k = 1, 2, \ldots, m$) трехмерных линейных систем

$$\begin{cases} a_k x_{2k-2}^2 + b_k x_{2k-2} + c_k = y_{2k-2}, \\ a_k x_{2k-1}^2 + b_k x_{2k-1} + c_k = y_{2k-1}, \\ a_k x_{2k}^2 + b_k x_{2k} + c_k = y_{2k}, \end{cases} \qquad (11.5)$$

соответствующим выставленным интерполяционным условиям.

Фактически, в рассмотренных случаях речь идет о последовательной линейной интерполяции (8.7) по перемещаемым вдоль отрезка $[a, b]$ парам соседних точек разбиения (11.1) и о последовательной квадратичной интерполяции (8.8) по тройкам таких точек.

Пример 11.1. Для функции $y = f(x)$, заданной таблицей

x	0	0.5	1	2	3	4	5
$f(x)$	1.5	0	0	2	2	1	2

выполним простейшие кусочно-линейное и кусочно-квадратичное интерполирования.

Осуществляя линейное интерполирование данной функции на каждом из элементарных промежутков, определяемых соседними числами верхней строки таблицы, получаем, что можно считать $f(x) \approx \varphi_1(x)$, где

$$\varphi_1(x) = \begin{cases} -3x + 1.5, & x \in [0, 0.5], \\ 0, & x \in [0.5, 1], \\ 2x - 2, & x \in [1, 2], \\ 2, & x \in [2, 3], \\ -x + 5, & x \in [3, 4], \\ x - 3, & x \in [4, 5]. \end{cases}$$

Квадратичное интерполирование по тройкам известных точек отрезков $[0, 1]$, $[1, 3]$ и $[3, 5]$ приводит к приближенному равенству $f(x) \approx \varphi_2(x)$, где

$$\varphi_2(x) = \begin{cases} 3x^2 - 4.5x + 1.5, & x \in [0, 1], \\ -x^2 + 5x - 4, & x \in [1, 3], \\ x^2 - 8x + 17, & x \in [3, 5]. \end{cases}$$

Графики функций $\varphi_1(x)$ и $\varphi_2(x)$ показаны на рис. 11.1 и 11.2.

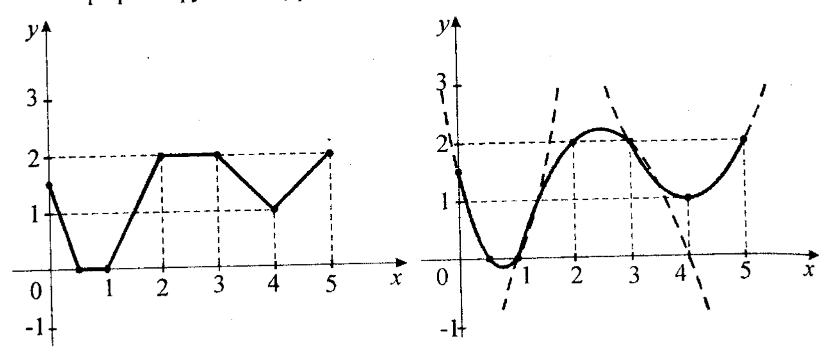


Рис. 11.1. *График функции* $y=\varphi_1(x)$ *Рис.* 11.2. *График функции* $y=\varphi_2(x)$

Кусочно-линейная аппроксимация таблично заданных функций, проводимая на основе метода наименьших квадратов, приводит к понятию линейного фильтра.

Предположим, что функция $f(x)$ задается «длинной» таблицей своих значений $f_i :\approx f(x_i)$ на системе равноотстоящих точек $x_i = x_0 + ih$, и пусть известно, что некоторые из значений f_i могут содержать достаточно большие случайные ошибки («выбросы»). Тогда, прежде чем производить какую-либо содержательную математическую обработку таких данных (например, полученных из эксперимента), целесообразно сначала произвести их ***сглаживание***, чтобы уменьшить роль выбросов. Рассмотрим вкратце один из простейших подходов к такому сглаживанию.

Если при кусочно-линейной и н т е р п о л я ц и и линейная функция на элементарном отрезке $[x_{i-1}, x_i]$ однозначно определялась по д в у м точкам из условия совпадения ее значений в них со значениями f_{i-1} и f_i данной функции $f(x)$, то здесь *на каждом из отрезков* $[x_{i-1}, x_{i+1}]$ *длиной* $2h$ *функция* $f(x)$ *будет подменяться наилучшим линейным среднеквадратическим приближением* $\varphi_i(x)$, *построенным по* т р е м *значениям* f_{i-1}, f_i *и* f_{i+1}, *и отфильтрованным значением в точке* x_i *будет считаться значение* $\varphi_i(x_i)$.

Итак, на первом этапе ставим задачу: для функции $f(x)$, представленной таблицей

x	x_0	x_1	$\dots$	x_n
$f(x)$	f_0	f_1	$\dots$	f_n

, (11.6)

где $x_i = x_0 + ih$, $i = 0, 1, \dots, n$, найти такую функцию $\varphi(x)$,

составленную из линейных функций

$$\varphi_i(x) := a_i + b_i(x - x_i)$$

(с перекрытием), чтобы

$$f(x) \approx \varphi_i(x) \qquad \forall\ x \in [x_{i-1}, x_{i+1}]$$

в смысле

$$\sum_{k=i-1}^{i+1} (f_k - \varphi_i(x_k))^2 = \min .$$

Эта знакомая по § 3.1 задача нахождения коэффициентов a_i, b_i методом наименьших квадратов решается непосредственно. Имеем:

$$(f_{i-1} - a_i + b_i h)^2 + (f_i - a_i)^2 + (f_{i+1} - a_i - b_i h)^2 = \min$$

$$\Leftrightarrow \begin{cases} (f_{i-1} - a_i + b_i h) + (f_i - a_i) + (f_{i+1} - a_i - b_i h) = 0, \\ (f_{i-1} - a_i + b_i h)h - (f_{i+1} - a_i - b_i h)h = 0. \end{cases}$$

Отсюда получаем $a_i = \frac{1}{3}(f_{i-1} + f_i + f_{i+1})$, $\quad b_i = \frac{1}{2h}(f_{i+1} - f_{i-1})$.

Следовательно, на каждом из промежутков $[x_{i-1}, x_{i+1}]$ функция $f(x)$ может быть подменена линейной функцией

$$\varphi_i(x) = \frac{1}{3}(f_{i-1} + f_i + f_{i+1}) + \frac{f_{i+1} - f_{i-1}}{2h}(x - x_i).$$

Варьированием i от 1 до $n-1$ для функции (11.6) таких линейных функций $\varphi_i(x)$ будет построено $n-1$; на каждый элементарный промежуток $[x_i, x_{i+1}]$ (кроме первого и последнего) их приходится по две.

Второй этап состоит в пересчете данной таблицы (11.6) заменой значения f_i на значение

$$\varphi_i(x_i) = \frac{1}{3}(f_{i-1} + f_i + f_{i+1})$$

при каждом $i = 1, 2, \dots, n-1$. Доопределив новую табличную функцию, например, значениями $\varphi_0(x_0) := f_0$ и $\varphi_n(x_n) := f_n$, получаем вместо (11.6) табличную зависимость

x	x_0	x_1	$\dots$	x_{n-1}	x_n
$\varphi(x)$	$\varphi_0(x_0)$	$\varphi_1(x_1)$	$\dots$	$\varphi_{n-1}(x_{n-1})$	$\varphi_n(x_n)$

,

в которой «в целом» сохраняется характер поведения исходной

функции $f(x)$ и уменьшена роль ее отдельных значений.

Описанная процедура называется ***осреднением по трем точкам*** и является простым частным случаем ***линейного фильтра***. Более глубокое осмысление роли линейных фильтров с помощью гармонического анализа можно найти в теории цифровой обработки сигналов; некоторые элементы такого анализа роли фильтров см., например, в [158, 187].

Отметим, что осреднение проводят и по большему числу точек, а иногда прибегают к двойной фильтрации табличных данных; правда, при глубокой фильтрации есть риск потерять полезную информацию о таблично заданной функции.

Пример 11.2. Функция $y = f(x)$, заданная таблицей

x	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
$f(x)$	0.0	0.1	0.2	0.3	1.0	0.4	0.5	0.5	0.1	0.6	0.7

,

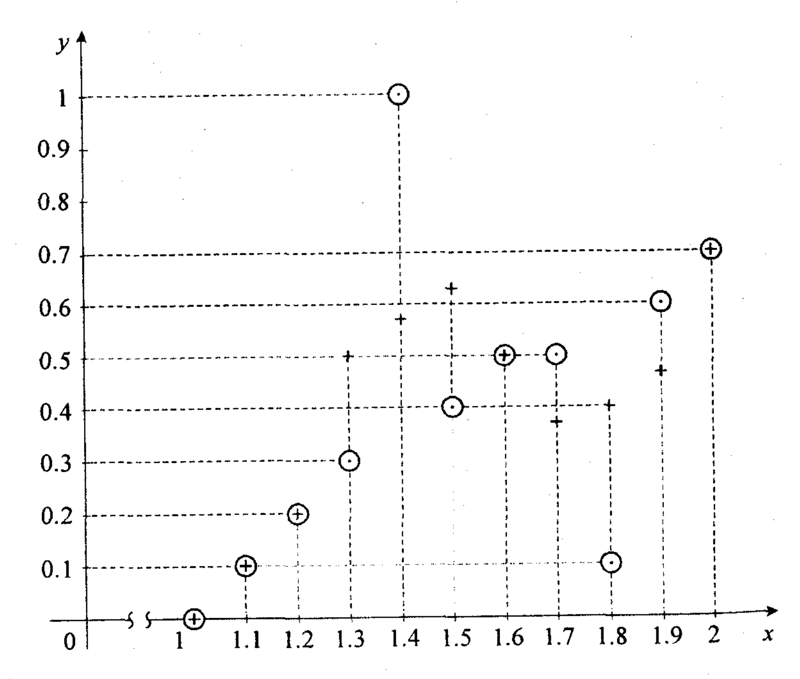


Рис. 11.3. Результат осреднения функции по трем точкам

представляет собой сеточную функцию $y = \ln x$ с систематическими ошибками округления в каждой точке сетки, не превосходящими 0.05, и двумя выбросами в точках $x = 1.4$ и $x = 1.8$. Осреднение по трем соседним точкам, т.е. замена каждого значения данной табличной функции средним

арифметическим трех ближайших значений, приводит к таблице

x	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
$\varphi(x)$	0.0	0.1	0.2	0.50	0.57	0.63	0.5	0.37	0.40	0.47	0.7

(где значения, подвергшиеся коррекции, записаны с двумя знаками после запятой). Эффект сглаживания, полученный таким простым фильтром, можно увидеть на рис. 11.3, где данные приближенные значения $f(x_i)$ помечены кружочками, а осредненные значения $\varphi(x_i)$ — крестиками.

11.2. ОПРЕДЕЛЕНИЕ СПЛАЙНА. ИНТЕРПОЛЯЦИОННЫЙ КУБИЧЕСКИЙ СПЛАЙН ДЕФЕКТА 1

Пусть на отрезке $[a, b]$ задана упорядоченная система несовпадающих точек x_k $(k = 0, 1, \ldots, n)$.

Определение 11.1. ***Сплайном*** *$S_m(x)$ называется определенная на $[a, b]$ функция, принадлежащая классу $C^l[a, b]$ l раз непрерывно дифференцируемых функций, такая, что на каждом промежутке $[x_{k-1}, x_k]$ $(k = 1, 2, \ldots, n)$ — это многочлен m-й степени. Разность $d := m - l$ между степенью сплайна m и показателем его гладкости l называется* ***дефектом*** *сплайна.*

Если сплайн $S_m(x)$ строится по некоторой функции $f(x)$ так, чтобы выполнялись условия $S_m(x_i) = f(x_i)$, то такой сплайн называется ***интерполяционным сплайном*** для функции $f(x)$; при этом ***узлы сплайна*** x_k, вообще говоря, могут не совпадать с узлами интерполяции x_i.

Тривиальные примеры интерполяционных сплайнов можно найти в предыдущем параграфе: кусочно-линейная функция $\varphi(x)$, определенная в (11.2) с параметрами a_k, b_k, удовлетворяющими условиям (11.3), очевидно, является интерполяционным сплайном степени 1 дефекта 1, а кусочно-квадратичная функция (11.4) при условиях (11.5) есть интерполяционный сплайн степени 2 дефекта 2.

Совпадение дефекта сплайна с его степенью обеспечивает просто непрерывность сплайна. Интерес представляет построение сплайнов с большей гладкостью, т. е. с малым дефектом. Такие сплайны являют собой дальнейшее совершенствование идеи кусочно-полиномиальной аппроксимации.

Рассмотрим наиболее известный и широко применяемый интерполяционный сплайн степени 3 дефекта 1. При этом будем исходить из предположения, что узлы сплайна

$$a = x_0,\ x_1,\ x_2,\ \ldots,\ x_{n-1},\ x_n = b \tag{11.7}$$

одновременно служат узлами интерполяции, т.е. в них известны значения функции $f_k := f(x_k)$, $k = 0, 1, \ldots, n$.

Определение 11.2. ***Кубическим сплайном дефекта 1****, интерполирующим на отрезке $[a, b]$ данную функцию $f(x)$, называется функция*

$$g(x) := \left\{ g_k(x) := a_k + b_k(x - x_k) + c_k(x - x_k)^2 + d_k(x - x_k)^3 \right.$$
$$\left. \text{при} \quad x \in [x_{k-1}, x_k] \right\}_{k=1}^{n}, \tag{11.8}$$

удовлетворяющая совокупности условий:

а) $g(x_k) = f_k$ *(условие интерполяции в узлах сплайна);*

б) $g(x) \in C^2[a, b]$ *(двойная непрерывная дифференцируемость);*

в) $g''(a) = g''(b) = 0$ *(краевые условия).*

Определенный таким образом сплайн называют еще ***естественным*** или ***чертежным сплайном****) и связано это со следующим обстоятельством. Желая провести плавную линию через заданные точки плоскости, чертежники фиксировали в этих точках гибкую упругую рейку, тогда под влиянием упругих сил она принимала нужную форму, обеспечивающую минимум потенциальной энергии. Несложно убедиться, что определяемая условиями а)–в) функция (11.8), представляющая собой кубический n-звенник с гладким сопряжением звеньев, служит математическим описанием такого чертежного приема.

С этой целью достаточно показать, что значение функционала $\Phi(g) := \int_a^b (g''(x))^2 dx$, характеризующее указанную величину потенциальной энергии закрепленной в $(n+1)$-й точке (11.7) упругой рейки, не превосходит величины $\Phi(f) := \int_a^b (f''(x))^2 dx$, соответствующей потенциальной энергии закрепленной в тех же точках рейки, но принимающей любую другую, отличную от $y = g(x)$ форму $y = f(x)$.

Более удобно это сделать, рассматривая $\Phi(f - g)$. Имеем:

$$\Phi(f-g) = \int_a^b (f''(x) - g''(x))^2 dx = \Phi(f) + \Phi(g) - 2\int_a^b f''(x) g''(x) dx =$$

(прибавим и вычтем величину $2\Phi(g)$)

$$= \Phi(f) - \Phi(g) - 2\int_a^b (f''(x) - g''(x)) g''(x) dx =$$

(применим правило интегрирования «по частям»)

$$= \Phi(f) - \Phi(g) - 2(f'(x) - g'(x)) g''(x)\Big|_a^b + 2\int_a^b (f'(x) - g'(x)) g'''(x) dx =$$

(при подстановке границ учтем краевые условия в), а к последнему интегралу применим свойство аддитивности по промежутку интегрирования)

$$= \Phi(f) - \Phi(g) + 2\sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} (f'(x) - g_k'(x)) g_k'''(x) dx =$$

$$= \Phi(f) - \Phi(g) + 12\sum_{k=1}^{n} d_k [f(x) - g_k(x)]_{x_{k-1}}^{x_k} =$$

(используем условия интерполяции, согласно которым должно быть $g(x_k) = f(x_k)$ и $g(x_{k-1}) = f(x_{k-1})$)

$$= \Phi(f) - \Phi(g).$$

Таким образом, $\Phi(g) = \Phi(f) - \Phi(f - g)$, откуда, в силу неотрицательности функционала Φ, следует $\Phi(g) \le \Phi(f)$.

Замечание 11.1. Краевые условия в определении 11.2 могут быть заменены на другие. Например, можно наложить дополнительные условия на первую производную функции $g(x)$ в точках a и b. В таком случае кубический сплайн (11.8), оставаясь интерполяционным дефекта 1, утрачивает свойство быть естественным.

Для построения по данной функции $f(x)$ интерполирующего ее сплайна (11.8) нужно найти $4n$ его коэффициентов a_k, b_k, c_k, d_k ($k = 1, 2, \ldots, n$). Чтобы понять, имеется ли для этого достаточное количество связей, расшифруем фигурирующие в определении 11.2 условия а)–в) через функции $g_k(x)$, составляющие $g(x)$, имея в виду, что в любом внутреннем узле должны

*) Английское слово *spline* переводится как (гибкая) планка, рейка, брусок.

совпадать значения двух соседних звеньев сплайна и двух их первых производных (см. наглядную структуру сплайна на рис. 11.4).

$$a \; g_1(x) \; g_2(x) \cdots g_{k-1}(x) \; g_k(x) \cdots g_n(x) \; b$$
$$x_0 \quad x_1 \quad x_2 \quad \cdots \quad x_{k-2} \quad x_{k-1} \quad x_k \quad \cdots \quad x_{n-1} \quad x_n$$

Рис. 11.4. *Расположение узлов и звеньев кубического сплайна* (11.8)

Имеем:

из условий интерполяции а):

$$g_1(x_0) = f_0, \quad g_k(x_k) = f_k \text{ при } k = 1, 2, \ldots, n;$$

из условий гладкой стыковки звеньев сплайна б):

$$\left.\begin{aligned} g_{k-1}(x_{k-1}) &= g_k(x_{k-1}), \\ g'_{k-1}(x_{k-1}) &= g'_k(x_{k-1}), \\ g''_{k-1}(x_{k-1}) &= g''_k(x_{k-1}) \end{aligned}\right\} \text{ при } k = 2, 3, \ldots, n;$$

из краевых условий в):

$$g''_1(x_0) = 0, \quad g''_n(x_n) = 0.$$

Как видим, условий оказалось $4n$ — ровно столько, сколько в записи сплайна (11.8) неизвестных коэффициентов. Подставляя сюда выражения функций

$$g_k(x) := a_k + b_k(x - x_k) + c_k(x - x_k)^2 + d_k(x - x_k)^3$$

и их производных

$$g'_k(x) = b_k + 2c_k(x - x_k) + 3d_k(x - x_k)^2 \tag{11.9}$$

и

$$g''_k(x) = 2c_k + 6d_k(x - x_k)$$

через коэффициенты a_k, b_k, c_k, d_k при указанных значениях k и полагая для краткости

$$h_k := x_k - x_{k-1}, \tag{11.10}$$

получаем детализированную систему связей

$$\begin{cases} a_1 - b_1h_1 + c_1h_1^2 - d_1h_1^3 = f_0, & (11.11) \\ a_k = f_k \quad \text{при } k = 1, 2, \ldots, n, & (11.12) \\ \left.\begin{aligned} a_{k-1} &= a_k - b_kh_k + c_kh_k^2 - d_kh_k^3, \\ b_{k-1} &= b_k - 2c_kh_k + 3d_kh_k^2, \\ c_{k-1} &= c_k - 3d_kh_k, \end{aligned}\right\} \text{ при } k = 2, 3, \ldots, n & \begin{aligned} (11.13) \\ (11.14) \\ (11.15) \end{aligned} \\ c_1 - 3d_1h_1 = 0, & (11.16) \\ c_n = 0. & (11.17) \end{cases}$$

Теперь ставим задачу выявления эффективного способа нахождения коэффициентов сплайна a_k, b_k, c_k, d_k ($k = 1, 2, \ldots, n$) из этой линейной относительно них системы (11.11)–(11.17). С этой целью будем исключать из системы неизвестные a_k, d_k, b_k и сводить все к решению системы относительно неизвестных c_k. При этом для упрощения записей используем уже применявшееся ранее (см. (8.39) в § 8.6) обозначение разделенной разности

$$f(x_{k-1}; x_k) := \frac{f_k - f_{k-1}}{h_k} \tag{11.18}$$

и, кроме того, введем фиктивный коэффициент

$$c_0 = 0. \tag{11.19}$$

Итак, согласно (11.12), коэффициенты a_k известны и равны f_k при любом $k \in \{1, 2, \ldots, n\}$. Подставляя их значения в равенства (11.11) и (11.13), приходим к равенству

$$b_kh_k - c_kh_k^2 + d_kh_k^3 = f_k - f_{k-1},$$

справедливому при $k = 1, 2, \ldots, n$, откуда с учетом обозначения (11.18) получаем выражение

$$b_k = f(x_{k-1}; x_k) + c_kh_k - d_kh_k^2 \quad (k = 1, 2, \ldots, n). \tag{11.20}$$

Далее, из (11.15) и (11.16), если учесть (11.19), можно однозначно выразить d_k через c_k:

$$d_k = \frac{c_k - c_{k-1}}{3h_k} \quad (k = 1, 2, \ldots, n). \tag{11.21}$$

С помощью (11.21) избавляемся от d_k в (11.20):

$$b_k = f(x_{k-1}; x_k) + \frac{2}{3}h_kc_k + \frac{1}{3}h_kc_{k-1} \quad (k = 1, 2, \ldots, n). \tag{11.22}$$

Теперь пользуемся связью (11.14), подставляя туда b_{k-1}, b_k из (11.22) и d_k из (11.21):

$$f(x_{k-2};\, x_{k-1}) + \frac{2}{3}h_{k-1}c_{k-1} + \frac{1}{3}h_{k-1}c_{k-2} =$$

$$= f(x_{k-1};\, x_k) + \frac{2}{3}h_k c_k + \frac{1}{3}h_k c_{k-1} - 2h_k c_k + 3h_k^2 \cdot \frac{c_k - c_{k-1}}{3h_k}.$$

После упрощения отсюда получаем относительно неизвестных c_k трехточечное разностное уравнение второго порядка

$$h_{k-1}c_{k-2} + 2(h_{k-1} + h_k)c_{k-1} + h_k c_k =$$

$$= 3f(x_{k-1};\, x_k) - 3f(x_{k-2};\, x_{k-1}), \quad (11.23)$$

где $k = 2, 3, \ldots, n$, $c_0 = 0$ (согласно (11.19)) и $c_n = 0$ (согласно (11.17)).

К такому же итогу можно прийти и иными путями [134, 138, 158, 167 и др.].

Так как $|2h_{k-1} + 2h_k|$ заведомо больше, чем $|h_{k-1}| + |h_k|$, то для разностного уравнения (11.23) (по другой терминологии, ленточной системы линейных алгебраических уравнений с трехдиагональной матрицей коэффициентов) выполняется достаточное условие однозначной разрешимости, т.е. существует единственный набор коэффициентов $c_1, c_2, \ldots, c_n$, удовлетворяющий (11.23). Найдя эти коэффициенты, все остальные коэффициенты сплайна (11.8) так же однозначно получаем по формулам (11.12), (11.22) и (11.21). Таким образом, справедлива следующая теорема.

Теорема 11.1. *При заданных в точках a и b краевых условиях ($g''(a) = g''(b) = 0$) по заданным в узлах*

$$a = x_0,\ x_1,\ x_2,\ \ldots,\ x_{n-1},\ x_n = b$$

значениям f_k ($k = 0, 1, \ldots, n-1$) функции $f(x)$ можно построить единственный интерполирующий ее кубический сплайн $g(x)$ дефекта 1.

Следует отметить высокую технологичность процесса сплайн-интерполирования.

Действительно, упомянутое выше диагональное преобладание в трехдиагональной матрице системы (11.23) обеспечивает корректность и устойчивость ***метода прогонки*** (см § 2.6 и [3, 44, 111, 153]. Применение этого метода для решения системы (11.23) сводится к вычислению прогоночных коэффициентов по формулам прямой прогонки

$$\delta_1 = -\frac{h_2}{2h_1 + 2h_2}, \quad \lambda_1 = \frac{3f(x_1;\, x_2) - 3f(x_0;\, x_1)}{2h_1 + 2h_2}, \quad (11.24)$$

$$\delta_{k-1} = -\frac{h_k}{2h_{k-1} + 2h_k + h_{k-1}\delta_{k-2}}, \quad (11.25)$$

$$\lambda_{k-1} = \frac{3f(x_{k-1};\, x_k) - 3f(x_{k-2};\, x_{k-1}) - h_{k-1}\lambda_{k-2}}{2h_{k-1} + 2h_k + h_{k-1}\delta_{k-2}} \quad (11.26)$$

при $k = 3, 4, \ldots, n$, а затем к получению искомых значений c_k обратной прогонкой по формуле

$$c_{k-1} = \delta_{k-1}c_k + \lambda_{k-1}, \quad (11.27)$$

полагая в ней $k = n, n-1, \ldots, 2$ и учитывая, что $c_n = 0$. После этого, как уже говорилось, остается подставить c_k в выражения d_k (11.21) и b_k (11.22), а в качестве a_k в (11.8) взять значения f_k. Все вычислительные затраты на построение n-звенного естественного сплайна составят, очевидно, $O(n)$ арифметических операций.

Факт и скорость сходимости сплайн-интерполяционного процесса характеризует следующее приводимое здесь без доказательства утверждение [1, 23, 117, 153, 167]

Теорема 11.2. *Пусть $g(x)$ — кубический сплайн дефекта 1, интерполирующий на системе узлов* (11.7) *отрезка $[a, b]$ четырежды непрерывно дифференцируемую на нем функцию $f(x)$. Тогда при любом фиксированном n найдется такая постоянная $C > 0$, что для любого $x \in [a, b]$*

$$|f(x) - g(x)| \le C \cdot \Delta^4, \quad (11.28)$$

где $\Delta := \max\limits_{1 \le k \le n} (x_k - x_{k-1})$.

Все расчетные формулы упрощаются в частном случае, когда сплайн $g(x)$ строится по системе р а в н о о т с т о я щ и х узлов. Выпишем всю совокупность формул в естественном для вычисления коэффициентов сплайна (11.8) порядке, заменяя в соответствующих равенствах (11.24)–(11.26), (11.22), (11.21) переменный шаг h_k (11.10) на постоянный шаг h. При этом вместо разделенных разностей (11.18) будут использовoваться к о н е ч н ы е разности интерполируемой функции $f(x)$ (см. (8.19) в § 8.4).

Имеем:

$$\delta_1 := -\frac{1}{4}, \quad \lambda_1 := \frac{3}{4h^2}\Delta^2 f_0;$$

при $k = 3, 4, \ldots, n$:

$$\delta_{k-1} = -\frac{1}{4+\delta_{k-2}}, \quad \lambda_{k-1} = \frac{\frac{3}{h^2}\Delta^2 f_{k-2} - \lambda_{k-2}}{4+\delta_{k-2}};$$

$c_n := 0$;

при $k = n, n-1, \ldots, 2$:

$$c_{k-1} = \delta_{k-1}c_k + \lambda_{k-1};$$

при $k = 1, 2, \ldots, n$ (с учетом $c_0 := 0$):

$$b_k = \frac{\Delta f_{k-1}}{h} + \frac{h}{3}(2c_k + c_{k-1}), \quad d_k = \frac{c_k - c_{k-1}}{3h}.$$

В результате при $x \in [x_{k-1}, x_k]$ значение $f(x)$ может быть подменено значением

$$g_k(x) = f_k + b_k(x - x_k) + c_k(x - x_k)^2 + d_k(x - x_k)^3$$

с найденными значениями коэффициентов b_k, c_k, d_k и погрешностью $O(h^4)$ в соответствии с оценкой (11.28).

Замечание 11.2. Если известно, что значения f_k приближаемой на $[a, b]$ функции $f(x)$ содержат случайные ошибки, то вместо интерполяционного сплайна применяют ***сглаживающий сплайн***. Сглаживающим сплайном, наиболее близким к естественному сплайну третьей степени, является такой кубический n-звенник $S_3(x)$, который минимизирует функционал

$$\Phi_1(u) := \int_a^b (u''(x))^2 dx + \sum_{k=0}^{n} p_k [u(x_k) - f_k]^2,$$

где $p_k > 0$ — некоторые весовые коэффициенты, от величины которых зависит степень сглаживания (или иначе, допустимое отклонение сглаживающего сплайна от интерполяционного в каждом конкретном узле) [24, 117 и др.]. Процедура построения сглаживающих сплайнов более сложна, и для получения коэффициентов, например, кубического сглаживающего сплайна $S_3(x)$ требуется решать линейные алгебраические системы с пятидиагональными матрицами [117, 134].

11.3. КВАДРАТИЧНЫЙ СПЛАЙН ДЕФЕКТА 1

Следуя [167] (см. также [73, 114 и др.]), будем строить такой интерполяционный сплайн второй степени, узлы которого не совпадают с узлами интерполяции, а перемежаются с ними (за исключением начального и последнего). А именно, пусть аппроксимируемая функция $f(x)$ на системе $n+1$ точек

$$a = x_0 < x_1 < x_2 < \ldots < x_n = b$$

— узлов интерполяции — принимает значения $f(x_k) = f_k$ $(k = 0, 1, \ldots, n)$, и пусть все внутренние ***узлы*** $\tilde{x}_k$ ***квадратичного сплайна*** $p(x)$ берутся точно посередине между соседними узлами интерполяции, т.е.

$$\tilde{x}_k = \frac{1}{2}(x_{k-1} + x_k),$$

а крайними узлами сплайна служат концы отрезка $[a, b]$:

$$\tilde{x}_0 := a\ (= x_0), \qquad \tilde{x}_{n+1} := b\ (= x_n).$$

Таким образом, на каждом из $n+1$ отрезков $[\tilde{x}_{k-1}, \tilde{x}_k]$, где $k = 1, 2, \ldots, n+1$ (рис. 11.5), сплайн $p(x)$ определяется своим ***звеном*** $p_k(x)$, которое будем задавать квадратичной функцией вида

$$p_k(x) := a_k + b_k(x - \tilde{x}_k) + c_k(x - \tilde{x}_k)^2. \tag{11.29}$$

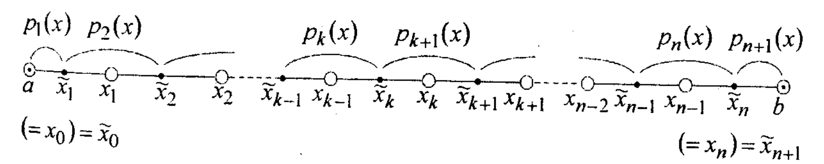


Рис. 11.5. Расположение узлов и звеньев квадратичного сплайна

Чтобы ***квадратичный сплайн***

$$p(x) := \{\, p_k(x), \quad x \in [\tilde{x}_{k-1}, \tilde{x}_k]\,\}_{k=1}^{n+1}$$

был интерполяционным для функции $f(x)$ и имел дефект 1, нужно удовлетворить условиям интерполяции во всех точках согласования функций $f(x)$ и $p(x)$:

$$p_{k+1}(x_k) = f_k \qquad \text{при } k = 0, 1, \ldots, n, \tag{11.30}$$

и условиям гладкой стыковки звеньев сплайна во всех его внутренних узлах:

$$\left.\begin{aligned} p_{k+1}(\widetilde{x}_k) &= p_k(\widetilde{x}_k) \\ p'_{k+1}(\widetilde{x}_k) &= p'_k(\widetilde{x}_k) \end{aligned}\right| \quad \text{при } k = 1, 2, \ldots, n. \tag{11.31}$$

В совокупности выписанные условия дают $3n+1$ уравнений относительно $3n+3$ неизвестных коэффициентов a_k, b_k, c_k, фигурирующих в выражениях функций $p_k(x)$ (11.29), составляющих $p(x)$. Оставшиеся свободными две связи можно реализовать по-разному; остановимся на задании краевых условий на вторую производную сплайна*):

$$p''_1(a) = A, \qquad p''_{n+1}(b) = B. \tag{11.32}$$

Продифференцировав дважды функцию (11.29), на основе условий (11.30)–(11.32) получаем следующую систему уравнений:

$$\begin{cases} a_{k+1} + b_{k+1}(x_k - \widetilde{x}_{k+1}) + c_{k+1}(x_k - \widetilde{x}_{k+1})^2 = f_k \quad (k=0,1,\ldots,n), \\ \left.\begin{aligned} &a_{k+1} + b_{k+1}(\widetilde{x}_k - \widetilde{x}_{k+1}) + c_{k+1}(\widetilde{x}_k - \widetilde{x}_{k+1})^2 = a_k, \\ &b_{k+1} + 2c_{k+1}(\widetilde{x}_k - \widetilde{x}_{k+1}) = b_k, \end{aligned}\right\} (k=1,2,\ldots,n) \\ 2c_1 = A, \quad 2c_{n+1} = B. \end{cases} \tag{11.33}$$

Для более удобного манипулирования с этой системой введем шаг интерполяции

$$h_k := x_k - x_{k-1}, \tag{11.34}$$

через который выразим фигурирующие в ней величины:

$$x_0 - \widetilde{x}_1 = -\frac{1}{2}h_1, \; x_k - \widetilde{x}_{k+1} = -\frac{1}{2}h_{k+1}, \; \widetilde{x}_k - \widetilde{x}_{k+1} = -\frac{1}{2}(h_k + h_{k+1})$$

при $k = 1, 2, \ldots, n-1$ и

$$x_n - \widetilde{x}_{n+1} = 0, \qquad \widetilde{x}_n - \widetilde{x}_{n+1} = -\frac{1}{2}h_n \quad \text{при } k = n.$$

*) В [167] показывается, что задание обеих дополнительных связей на одном конце промежутка $[a, b]$ приводит к численно неустойчивому процессу сплайн-интерполяции.

С обозначениями (11.34) система (11.33) приобретает вид

$$\begin{cases} a_{k+1} - \frac{1}{2}h_{k+1}b_{k+1} + \frac{1}{4}h_{k+1}^2 c_{k+1} = f_k \quad (k = 0, 1, \ldots, n-1), \\ a_{n+1} = f_n, \\ a_{k+1} - \frac{1}{2}(h_k + h_{k+1})b_{k+1} + \frac{1}{4}(h_k + h_{k+1})^2 c_{k+1} = a_k \\ \qquad\qquad (k = 1, 2, \ldots, n-1), \\ a_{n+1} - \frac{1}{2}h_n b_{n+1} + \frac{1}{4}h_n^2 c_{n+1} = a_n, \\ b_{k+1} - (h_k + h_{k+1})c_{k+1} = b_k \quad (k = 1, 2, \ldots, n-1), \\ b_{n+1} - h_n c_{n+1} = b_n, \\ 2c_1 = A, \quad 2c_{n+1} = B. \end{cases} \tag{11.35}$$

Результат преобразований этой системы будет приведен ниже (см. формулы (11.52)–(11.55)), а сейчас рассмотрим упрощенный случай, когда параболическая сплайн-интерполяция совершается по системе равноотстоящих узлов.

Будем считать, что шаг интерполяции h_k постоянен и равен $2h$, т.е. пусть в (11.34)

$$h_k = 2h \qquad \forall\, k \in \{1, 2, \ldots, n\}.$$

Тогда коэффициенты a_k, b_k, c_k звеньев сплайна (11.29) могут быть найдены из следующего частного случая системы (11.35):

$$\begin{cases} a_{k+1} - hb_{k+1} + h^2 c_{k+1} = f_k \quad (k = 0, 1, \ldots, n-1), & (11.36) \\ a_{n+1} = f_n, & (11.37) \\ a_{k+1} - 2hb_{k+1} + 4h^2 c_{k+1} = a_k \quad (k = 1, 2, \ldots, n-1), & (11.38) \\ a_{n+1} - hb_{n+1} + h^2 c_{n+1} = a_n, & (11.39) \\ b_{k+1} - 4hc_{k+1} = b_k \quad (k = 1, 2, \ldots, n-1), & (11.40) \\ b_{n+1} - 2hc_{n+1} = b_n, & (11.41) \\ 2c_1 = A, \quad 2c_{n+1} = B. & (11.42) \end{cases}$$

Как и в предыдущем параграфе, будем исключать из уравнений системы неизвестные a_k, b_k и сводить всё к связям между неизвестными c_k.

Так как из (11.36)

$$a_{k+1} = hb_{k+1} - h^2 c_{k+1} + f_k \quad \text{при} \quad k = 0, 1, \dots, n-1,$$

и следовательно,

$$a_k = hb_k - h^2 c_k + f_{k-1} \quad \text{при} \quad k = 1, 2, \dots, n, \tag{11.43}$$

то, воспользовавшись этим в равенстве (11.38), имеем

$$hb_{k+1} - h^2 c_{k+1} + f_k - hb_k + h^2 c_k - f_{k-1} - 2hb_{k+1} + 4h^2 c_{k+1} = 0,$$

т.е. при любых $k \in \{1, 2, \dots, n-1\}$ справедливо равенство

$$b_k + b_{k+1} = hc_k + 3hc_{k+1} + \frac{\Delta f_{k-1}}{h} \tag{11.44}$$

(напомним, что $\Delta f_{k-1} := f_k - f_{k-1}$ — конечная разность первого порядка функции $f(x)$). Рассматривая (11.44) совместно с равенством (11.40), получаем

$$2b_k = hc_k - hc_{k+1} + \frac{\Delta f_{k-1}}{h} \quad \text{при} \quad k = 1, 2, \dots, n-1, \tag{11.45}$$

и значит,

$$2b_{k+1} = hc_{k+1} - hc_{k+2} + \frac{\Delta f_k}{h} \quad \text{при} \quad k = 0, 1, \dots, n-2. \tag{11.46}$$

Подставим (11.45) и (11.46) в равенство (11.44) (предварительно умноженное на 2):

$$hc_k - hc_{k+1} + \frac{\Delta f_{k-1}}{h} + hc_{k+1} - hc_{k+2} + \frac{\Delta f_k}{h} = 2hc_k + 6hc_{k+1} + \frac{2\Delta f_{k-1}}{h}.$$

Отсюда после приведения подобных членов, деления на h и введения конечной разности второго порядка $\Delta^2 f_{k-1} := \Delta f_k - \Delta f_{k-1}$ приходим к трехточечному рекуррентному равенству

$$c_k + 6c_{k+1} + c_{k+2} = \frac{\Delta^2 f_{k-1}}{h^2}, \tag{11.47}$$

связывающему неизвестные величины c_k при $k = 1, 2, \dots, n-2$.

При последнем из допустимых в (11.47) значений $k = n-2$ получается уравнение, содержащее неизвестные c_{n-2}, c_{n-1} и c_n. Чтобы вывести уравнение, связывающее c_{n-1}, c_n и c_{n+1}, воспользуемся незадействованными еще соотношениями (11.37), (11.39) и (11.41).

Сначала подставим $a_{n+1} = f_n$ и $a_n = hb_n - h^2 c_n + f_{n-1}$ (см. (11.36) при $k = n-1$) в (11.39), в результате чего получим

$$b_{n+1} + b_n = hc_n + hc_{n+1} + \frac{\Delta f_{n-1}}{h}.$$

Рассматривая это равенство совместно с (11.41), т.е. с $b_{n+1} - b_n = 2hc_{n+1}$, находим выражение

$$2b_n = hc_n - hc_{n+1} + \frac{\Delta f_{n-1}}{h}. \tag{11.48}$$

Теперь обратимся к равенству (11.45), согласно которому при $k = n-1$ имеем

$$2b_{n-1} = hc_{n-1} - hc_n + \frac{\Delta f_{n-2}}{h}. \tag{11.49}$$

Но, в силу (11.44) при $k = n-1$, должно быть

$$2b_n + 2b_{n-1} = 2hc_{n-1} + 6hc_n + \frac{2\Delta f_{n-2}}{h}.$$

Складывая равенства (11.48) и (11.49) и сравнивая полученное с последним равенством, после упрощений приходим к равенству

$$c_{n-1} + 6c_n + c_{n+1} = \frac{\Delta^2 f_{n-2}}{h^2}, \tag{11.50}$$

имеющему абсолютно ту же структуру, что и (11.47); иначе говоря, (11.50) — это (11.47) при $k = n-1$.

Итак, собирая воедино формулы, нужные для цельного описания процедуры построения интерполяционного квадратичного сплайна

$$p(x) := \{ p_k(x), \quad x \in [\tilde{x}_{k-1}, \tilde{x}_k] \}_{k=1}^{n+1}$$

дефекта 1 по системе равноотстоящих с шагом $2h$ узлов интерполяции x_k и узлов сплайна $\tilde{x}_k = x_{k-1} + h$, имеем следующее.

Решая трехдиагональную систему

$$\begin{cases} 2c_1 = A, \\ 6c_2 + c_3 = \dfrac{\Delta^2 f_0}{h^2} - \dfrac{A}{2}, \\ c_2 + 6c_3 + c_4 = \dfrac{\Delta^2 f_1}{h^2}, \\ \dots\dots\dots\dots\dots \\ c_{n-2} + 6c_{n-1} + c_n = \dfrac{\Delta^2 f_{n-3}}{h^2}, \\ c_{n-1} + 6c_n = \dfrac{\Delta^2 f_{n-2}}{h^2} - \dfrac{B}{2}, \\ 2c_{n+1} = B \end{cases} \qquad (11.51)$$

(например, прогонкой, которая здесь заведомо корректна и устойчива), находим коэффициенты c_k $(k = 1, 2, \dots, n+1)$ звеньев сплайна $p_k(x)$ (11.29); обоснованием этой системы служат формулы (11.42), (11.47), (11.50). По известным c_k далее вычисляем

$$b_k = \frac{h}{2}(c_k - c_{k+1}) + \frac{1}{2}\frac{\Delta f_{k-1}}{h} \quad \text{при} \quad k = 1, 2, \dots, n$$

в соответствии с (11.45) и (11.48); после этого находим

$$b_{n+1} = b_n + Bh$$

на основе (11.41) с учетом (11.42). Последними вычисляются коэффициенты a_k (см. (11.43) и (11.37)):

$$a_k = h(b_k - hc_k) + f_{k-1} \quad \text{при } k = 1, 2, \dots, n,$$
$$a_{n+1} = f_n.$$

Замечание 11.3. Если сравнить систему (11.51) с аналогичной системой, получающейся из (11.23) при $h_k = h$ в случае построения кубического сплайна по равноотстоящим узлам, можно заметить, что в системе (11.51) более ярко выражено диагональное преобладание; это позволяет надеяться на лучшую численную устойчивость процедуры нахождения коэффициентов квадратичного сплайна.

Вывод формул для получения коэффициентов c_k, b_k, a_k в общем случае переменного шага интерполяции h_k из совокупности условий (11.35) не вызывает затруднений и производится

аналогично продемонстрированному выше выводу. Итогом является следующая ленточная система (где $f(x_{k-1}; x_k; x_{k+1})$ — разделенные разности второго порядка, построенные по значениям f_{k-1}, f_k, f_{k+1}):

$$\begin{cases} 2c_1 = A, \\ 3c_2 + \dfrac{h_2}{h_1 + h_2}c_3 = 4f(x_0; x_1; x_2) - \dfrac{Ah_1}{2h_1 + 2h_2}, \\ \dfrac{h_2}{h_2 + h_3}c_2 + 3c_3 + \dfrac{h_3}{h_2 + h_3}c_4 = 4f(x_1; x_2; x_3), \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ \dfrac{h_{n-2}}{h_{n-2} + h_{n-1}}c_{n-2} + 3c_{n-1} + \dfrac{h_{n-1}}{h_{n-2} + h_{n-1}}c_n = 4f(x_{n-3}; x_{n-2}; x_{n-1}), \\ \dfrac{h_{n-1}}{h_{n-1} + h_n}c_{n-1} + 3c_n = 4f(x_{n-2}; x_{n-1}; x_n) - \dfrac{Bh_n}{2h_{n-1} + 2h_n}, \\ 2c_{n+1} = B\ . \end{cases} \qquad (11.52)$$

Решив ее относительно величин c_k, служащих коэффициентами при вторых степенях звеньев сплайна $p_k(x)$, далее вычисляем коэффициенты b_k при первых степенях, полагая $k = 1, 2, \dots, n$ в равенстве

$$b_k = \frac{1}{4}h_k(c_k - c_{k+1}) + f(x_{k-1}; x_k), \qquad (11.53)$$

а затем

$$b_{n+1} = b_n + \frac{1}{2}Bh_n, \qquad (11.54)$$

и, наконец, свободные члены a_k квадратных трехчленов $p_k(x)$, составляющих сплайн $p(x)$, находим с помощью равенств

$$a_k = \frac{1}{2}h_k\left(b_k - \frac{1}{2}h_k c_k\right) + f_{k-1}, \qquad (11.55)$$

где $k = 1, 2, \dots, n$ и $a_{n+1} = f_n$.

Проведенные выкладки и их результаты можно рассматривать как доказательство существования и единственности квадратичного сплайна дефекта 1, интерполирующего на отрезке $[a, b]$ заданную на нем своими значениями f_k функцию $f(x)$ при заданных краевых условиях (11.32). Ряд утверждений о

сходимости таких сплайнов к функции $f(x)$ на $[a, b]$ с тем или иным порядком относительно $h \to 0$ в равномерном и $\Delta := \max(x_k - x_{k-1}) \to 0$ в неравномерном случаях, в зависимости от требуемой гладкости $f(x)$, можно найти в [167], причем в этих утверждениях даются оценки близости в точках $x \in [a, b]$ не только $f(x)$ и $p(x)$, но и их производных*). В простейшем виде утверждения подобного рода качественно отражает следующая теорема.

Теорема 11.3. *Пусть функция $f(x)$ дважды непрерывно дифференцируема на отрезке $[a, b]$ и пусть $p(x)$ — квадратичный сплайн, определяемый условиями* (11.30)–(11.32). *Тогда при каждом фиксированном n найдутся такие положительные постоянные c_0, c_1, c_2, что при любом $x \in [a, b]$*

$$|f(x) - p(x)| \le c_0\Delta^2,$$

$$\left|f^{(i)}(x) - p^{(i)}(x)\right| \le c_i\Delta^{2-i} \qquad (i = 1, 2),$$

где $\Delta := \max\limits_{k\in\{1, \ldots, n\}} (x_k - x_{k-1})$.

При этом теорема сохраняет силу и в случаях, когда краевые условия (11.32) на вторую производную $p(x)$ подменяются соответствующими условиями на первую производную

$$p_1'(a) = A, \qquad p_{n+1}'(b) = B. \qquad (11.56)$$

или периодическими краевыми условиями**)

$$p_1^{(i)}(a) = p_{n+1}^{(i)}(b), \quad \text{где} \quad i = 1, 2.$$

Замечание 11.4. При построении интерполяционного квадратичного сплайна дефекта 1, соответствующего требованиям (11.30), (11.31) и (11.56), более удобным может оказаться подход, при котором вычисление коэффициентов a_k, b_k, c_k сводится к применению метода прогонки для нахождения значений b_k (а не c_k) из аналогичной (11.52) системы с последующей подстановкой их в формулы для получения a_k и c_k.

---

*) Отметим кстати, что существуют оценки близости производных функции $f(x)$ и кубического сплайна $g(x)$, подобные утверждаемым теоремой 11.3 для квадратичных сплайнов $p(x)$.

**) В последнем случае точка a не должна быть узлом сплайна.

11.4. БАЗИСНЫЕ СПЛАЙНЫ

Без сомнений, можно построить ряд других сплайнов, подобных рассмотренным в предыдущих параграфах кубическим и квадратичным сплайнам, которые обладали бы теми или иными заданными свойствами, связанными со свойствами интерполируемых ими на отрезке $[a, b]$ функций $f(x)$. Среди таких сплайнов более высоких степеней стоит особо отметить, пожалуй, лишь сплайны степени 5 дефекта 3, играющие важную роль в приложениях. Существует достаточно общая и глубокая теория сплайнов, изучающая их свойства, представления, сходимость и т. п. [1, 91, 114, 167 и др.]. Имеется много разработок в области практических применений сплайнов [18, 24, 73, 82, 134 и др.].

Остановимся на нескольких специфических конструкциях кусочно-полиномиальных функций, тоже именующихся сплайнами (с некоторыми уточняющими их особенности прилагательными) и определенным образом связанные с обычными сплайнами.

Определение 11.3 [167]. ***Базисным сплайном*** *или, короче,* ***B-сплайном*** *степени $m-1$ (дефекта 1) относительно* ***узлов*** *$x_k < x_{k+1} < \ldots < x_{k+m}$ называется функция*

$$B_{m-1,k}(x) = B_{m-1}(x_k;\ x_{k+1};\ \ldots;\ x_{k+m};\ x) :=$$

$$= m\sum_{i=k}^{k+m} \frac{(x_i - x)_+^{m-1}}{\Pi'_{m,k}(x_i)}, \qquad (11.57)$$

где

$$(x_i - x)_+^{m-1} := [\max\{0; (x_i - x)\}]^{m-1}, \qquad (11.58)$$

$$\Pi_{m,k}(x) := (x - x_k)(x - x_{k+1})\ldots(x - x_{k+m}). \qquad (11.59)$$

Посмотрим, что представляют собой B-сплайны от нулевой до третьей степени в более простом, но и более употребительном случае равноотстоящих узлов B-сплайна, т.е. когда

$$x_{k+i} = x_k + ih, \qquad \text{где} \quad i = 0, 1, \ldots, m.$$

Сопоставляя обозначение (11.59) с встречавшимся ранее аналогичным обозначением (8.10), без труда «расшифровываем» знаменатели дробей в (11.57): $\Pi'_{m,k}(x_i)$ — это произведение разностей между узлом x_i и всеми остальными узлами от x_k до x_{k+m}.

Числители этих дробей в соответствии с (11.58) есть

$$(x_i - x)_+^{m-1} := \begin{cases} (x_i - x)^{m-1} & \text{при } x \le x_i\,, \\ 0 & \text{при } x \ge x_i\,. \end{cases} \tag{11.60}$$

Дальнейшую детализацию выражений B-сплайнов (11.57) будем производить, придавая параметру m значения 1, 2, 3, 4.

1) Пусть $m = 1$. Тогда из (11.57) имеем (считая, что $0^0 := 0$):

$$B_{0,k}(x) = B_0(x_k;\, x_{k+1};\, x) = \frac{(x_k - x)_+^0}{x_k - x_{k+1}} + \frac{(x_{k+1} - x)_+^0}{x_{k+1} - x_k} =$$

$$= \begin{cases} -\dfrac{1}{h} + \dfrac{1}{h} = 0 & \text{при } x < x_k\,, \\ 0 + \dfrac{1}{h} = \dfrac{1}{h} & \text{при } x_k \le x < x_{k+1}, \\ 0 + 0 = 0 & \text{при } x \ge x_{k+1}. \end{cases} \tag{11.61}$$

Таким образом, ***В-сплайн нулевой степени*** представляет собой функцию-«ступеньку» (рис. 11.6)*).

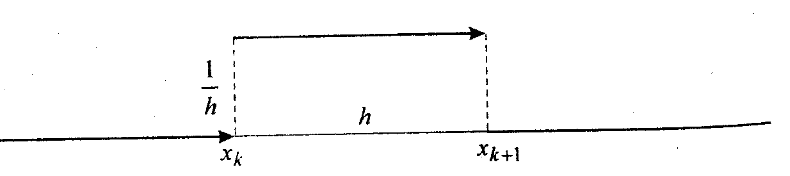


Рис. 11.6. *В-сплайн нулевой степени*

2) При $m = 2$ из (11.57) следует

$$B_{1,k}(x) = B_1(x_k;\, x_{k+1};\, x_{k+2};\, x) = 2\left[\frac{(x_k - x)_+^1}{(x_k - x_{k+1})(x_k - x_{k+2})} + \right.$$

*) Это единственный из базисных сплайнов (11.57), изменяющий свои значения скачком. К какому из узлов x_k или x_{k-1} отнести ненулевое значение $\frac{1}{h}$ — вопрос договоренности. То же можно сказать об $(m-1)$-х производных B-сплайнов $(m-1)$-й степени.

$$\left. + \frac{(x_{k+1} - x)_+^1}{(x_{k+1} - x_k)(x_{k+1} - x_{k+2})} + \frac{(x_{k+2} - x)_+^1}{(x_{k+2} - x_k)(x_{k+2} - x_{k+1})}\right] =$$

$$= \frac{1}{h^2}\left[(x_k - x)_+ - 2(x_{k+1} - x)_+ + (x_{k+2} - x)_+\right].$$

В соответствии с (11.60) последнее можно представить в виде

$$\begin{cases} \dfrac{1}{h^2}(x_k - x - 2x_{k+1} + 2x + x_{k+2} - x)\,, & \text{если } x \le x_k\,, \\ \dfrac{1}{h^2}(0 - 2x_{k+1} + 2x + x_{k+2} - x)\,, & \text{если } x_k \le x \le x_{k+1}\,, \\ \dfrac{1}{h^2}(0 - 0 + x_{k+2} - x)\,, & \text{если } x_{k+1} \le x \le x_{k+2}\,, \\ 0\,, & \text{если } x \ge x_{k+2}\,. \end{cases}$$

Простыми упрощениями получаем окончательное выражение ***линейного В-сплайна***:

$$B_{1,k}(x) = B_1(x_k;\, x_{k+1};\, x_{k+2};\, x) =$$

$$= \begin{cases} 0 & \text{при } x \le x_k\,, \\ \dfrac{1}{h} + \dfrac{x - x_{k+1}}{h^2} & \text{при } x_k \le x \le x_{k+1}\,, \\ \dfrac{1}{h} - \dfrac{x - x_{k+1}}{h^2} & \text{при } x_{k+1} \le x \le x_{k+2}\,, \\ 0 & \text{при } x \ge x_{k+2}\,. \end{cases} \tag{11.62}$$

Нетрудно представить геометрическое изображение линейного B-сплайна (рис. 11.7) — это так называемая функция-«крышка» (по-другому, функция-«шапочка» [31]).

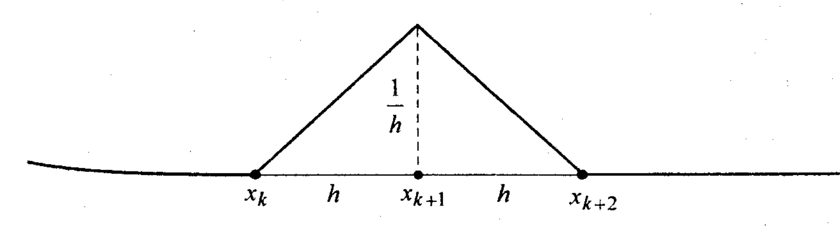


Рис. 11.7. *Линейный В-сплайн*

3) Полагая в (11.57) $m=3$ и выполняя несложные преобразования, выясняем, как выглядит ***квадратичный B-сплайн***:

$$B_{2,k}(x)=B_2(x_k;\,x_{k+1};\,x_{k+2};\,x_{k+3};\,x)=$$

$$=\frac{1}{h^3}\left[-\frac{1}{2}(x_k-x)_+^2+\frac{3}{2}(x_{k+1}-x)_+^2-\frac{3}{2}(x_{k+2}-x)_+^2+\frac{1}{2}(x_{k+3}-x)_+^2\right]=$$

$$=\begin{cases} 0 & \text{при } x\le x_k, \\ \frac{1}{2h}-\frac{1}{h^2}(x_{k+1}-x)+\frac{1}{2h^3}(x_{k+1}-x)^2 & \text{при } x_k\le x\le x_{k+1}, \\ \frac{1}{2h}+\frac{1}{h^2}(x_{k+2}-x)+\frac{1}{h^3}(x_{k+2}-x)^2 & \text{при } x_{k+1}\le x\le x_{k+2}, \\ \frac{1}{2h^3}(x_{k+3}-x)^2 & \text{при } x_{k+2}\le x\le x_{k+3}, \\ 0 & \text{при } x\ge x_{k+3}. \end{cases} \quad (11.63)$$

График такого сплайна изображен на рис. 11.8. Для его построения предварительно находим

$$B'_{2,k}(x)=\frac{1}{h^3}\begin{cases} h-(x_{k+1}-x), & x_k\le x\le x_{k+1}, \\ -h+2(x_{k+2}-x), & x_{k+1}\le x\le x_{k+2}, \\ -(x_{k+3}-x), & x_{k+2}\le x\le x_{k+3}, \end{cases}$$

откуда следует, что производная функции $B_{2,k}(x)$ непрерывна во всех ее узлах и что $\max B_{2,k}(x)=\frac{3}{4h}$ в точке $x=x_{k+1}+\frac{h}{2}=$ $=x_{k+2}-\frac{h}{2}$. Из выражения второй производной

$$B''_{2,k}(x)=\frac{1}{h^3}\begin{cases} 1, & x_k\le x<x_{k+1}, \\ -2, & x_{k+1}\le x<x_{k+2}, \\ 1, & x_{k+2}\le x<x_{k+3} \end{cases}$$

видим, что выпуклость функции $B_{2,k}(x)$ изменяется в узлах скачком.

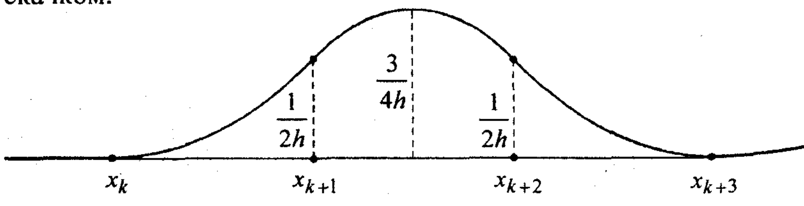


Рис. 11.8. *Квадратичный B-сплайн*

4) При $m=4$ аналогичным образом нетрудно получить из (11.57) ***кубический B-сплайн***

$$B_{3,k}(x)=\frac{1}{6h^4}\Big[(x_k-x)_+^3-4(x_{k+1}-x)_+^3+6(x_{k+2}-x)_+^3-$$
$$-4(x_{k+3}-x)_+^3+(x_{k+4}-x)_+^3\Big]=$$

$$=\begin{cases} 0, & x\le x_k, \\ \frac{1}{6h^4}(x-x_k)^3, & x_k\le x\le x_{k+1}, \\ \frac{1}{6h}+\frac{1}{2h^2}(x-x_{k+1})+\frac{1}{2h^3}(x-x_{k+1})^2- \\ \quad -\frac{1}{2h^4}(x-x_{k+1})^3, & x_{k+1}\le x\le x_{k+2}, \\ \frac{1}{6h}+\frac{1}{2h^2}(x_{k+3}-x)+\frac{1}{2h^3}(x_{k+3}-x)^2- \\ \quad -\frac{1}{2h^4}(x_{k+3}-x)^3, & x_{k+2}\le x\le x_{k+3}, \\ \frac{1}{6h^4}(x_{k+4}-x)^3, & x_{k+3}\le x\le x_{k+4}, \\ 0, & x\ge x_{k+4}. \end{cases}$$

Так же легко убедиться, что базисный сплайн третьей степени имеет непрерывными не только первые производные, как предыдущий, но и вторые. График этого сплайна представлен на рис. 11.9*).

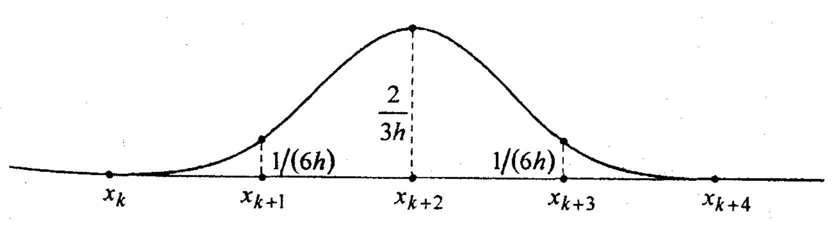


Рис. 11.9. *Кубический B-сплайн*

Являясь линейно независимыми функциями, B-сплайны определенной степени образуют базис в соответствующих

*) Выведенный здесь из общей формулы кубический B-сплайн отличается от аналогичного сплайна, например, из [138] множителем $\frac{2}{3h}$.

функциональных пространствах, что оправдывает их название и говорит о возможности представления через B-сплайны других функций этих пространств. Интуитивно ясно, что любая кусочно-постоянная функция на отрезке, составленном из равных промежутков длины h, может быть единственным образом представлена линейной комбинацией B-сплайнов нулевой степени (11.61), любая кусочно-линейная — B-сплайнов первой степени (11.62), и т.д. Эти факты могут быть четко сформулированы и строго обоснованы. Например, для изучавшегося в § 11.2 естественного сплайна справедливо следующее утверждение [138].

Теорема 11.4. *Пусть* $g(x)$ *— кубический сплайн* (11.8) *дефекта* 1, *построенный по системе равноотстоящих узлов* $x_k = x_0 + kh$ $(k = 0, 1, \ldots, n)$. *Тогда найдутся такие постоянные* α_0, $\alpha_1, \ldots, \alpha_{n+2}$, *что*

$$g(x) = \sum_{i=0}^{n+2} \alpha_i B_{3,i-3}(x). \qquad (11.64)$$

Заметим, что для построения функций $B_{3,-3}(x)$, $B_{3,-2}(x)$, $B_{3,-1}(x)$ и $B_{3,n-3}(x)$, $B_{3,n-2}(x)$, $B_{3,n-1}(x)$ в (11.64) требуется введение вспомогательных точек сетки (за пределами отрезка $[a,b]$) x_{-3}, x_{-2}, x_{-1} и x_{n+1}, x_{n+2}, x_{n+3} соответственно.

Приближение функций линейными комбинациями базисных сплайнов позволяет запоминать лишь коэффициенты α_i этих комбинаций. Вычисление же их, в свою очередь, может осуществляться достаточно эффективно, в силу очевидной ленточной структуры матриц линейных систем, к которым сводится нахождение α_i приравниванием левых и правых частей равенств типа (11.64) в узлах сплайнов $x = x_k$.

B-сплайны играют огромную роль при построении численно-аналитических методов решения дифференциальных и интегральных уравнений. В частности, линейные B-сплайны (11.62) лежат в основе весьма популярного метода конечных элементов (см. гл. 17).

11.5. ЭРМИТОВЫ (ЛОКАЛЬНЫЕ) СПЛАЙНЫ

Предположим, что у аппроксимируемой функции $y = f(x)$ в точках $x_0, x_1, \ldots, x_n \in [a, b]$ известны не только значения $y_0, y_1, \ldots, y_n$, но и значения производных до m-й включительно,

т.е. функция $y = f(x)$ задается таблицей

x	y	y'	$\ldots$	$y^{(m)}$
x_0	y_0	y'_0	$\ldots$	$y_0^{(m)}$
x_1	y_1	y'_1	$\ldots$	$y_1^{(m)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	y_n	y'_n	$\ldots$	$y_n^{(m)}$

(11.65)

Как известно (см. § 8.8), по этой информации об $f(x)$ можно построить единственный интерполяционный многочлен Эрмита степени $(n+1)(m+1)-1$. Даже при сравнительно небольших значениях n и m эта степень может оказаться неоправданно высокой. Здесь имеется в виду как сложность построения и эффективного использования многочленов Эрмита высоких степеней, так и повышенные требования к гладкости функций, для которых они должны служить приближениями.

Применение в таких случаях обычных интерполяционных сплайнов высоких степеней и малых дефектов дает нужную гладкость интерполируемой функции, т. е. самого сплайна, но не обеспечивает согласования его производных с данными производными функции $f(x)$.

Напрашивается сведение воедино этих двух способов аппроксимации функций, т. е. образование гибрида интерполяционного сплайна и интерполяционного многочлена Эрмита. Такие гибриды называют ***эрмитовыми сплайнами***.

Глядя на таблицу (11.65), в соответствии с полученным из § 8.8 представлением об эрмитовой интерполяции, легко вообразить, что по данным в точках x_0 и x_1 значениям функции $f(x)$ и ее m производных можно построить единственный многочлен Эрмита $H^1_{2m+1}(x)$ степени $2m+1$, по данным в точках x_1 и x_2 — многочлен такой же степени $H^2_{2m+1}(x)$, и т. д. Поскольку каждый из этих многочленов — звеньев эрмитова сплайна

$$S_{2m+1}(x) := \left\{ H^i_{2m+1}(x),\ x \in [x_{i-1}, x_i] \right\}_{i=1}^{n} \qquad (11.66)$$

— строится независимо от остальных, то такие сплайны называются также ***локальными сплайнами***.

Свойство локальности эрмитова сплайна дает определенные преимущества при его построении, особенно в случае большого числа n элементарных промежутков $[x_{i-1}, x_i]$, а также позволяет составлять эрмитов сплайн из звеньев разных степеней, если на каких-то участках промежутка интерполяции $[a, b]$ функция $f(x)$ представлена производными одного порядка, а на

других — другого порядка. При этом заметим, что в точках стыковки таких участков допустимо использование односторонних производных.

Посмотрим, что собой представляет **эрмитов сплайн**, например, **третьей степени** (эрмитов сплайн первой степени, по сути, определен в § 11.1 формулами (11.2)–(11.3)).

Считая, что на i-м элементарном промежутке функция $y = f(x)$ задается таблицей

x	y	y'
x_{i-1}	y_{i-1}	y'_{i-1}
x_i	y_i	y'_i

,

полагаем в (11.66) $m = 1$, и i-е звено сплайна $S_3(x)$ ищем в виде*)

$$H_3^i(x) = a_0^i + a_1^i(x - x_{i-1}) + a_2^i(x - x_{i-1})^2 + a_3^i(x - x_{i-1})^3. \quad (11.67)$$

Два уравнения для нахождения коэффициентов многочлена $H_3^i(x)$ получаем из интерполяционных условий $H_3^i(x_{i-1}) = y_{i-1}$, $H_3^i(x_i) = y_i$, а именно:

$$a_0^i = y_{i-1}, \quad (11.68)$$

$$a_0^i + a_1^i h_i + a_2^i h_i^2 + a_3^i h_i^3 = y_i, \quad (11.69)$$

где $h_i := x_i - x_{i-1}$. Далее дифференцируем (11.67):

$$\left(H_3^i\right)'(x) = a_1^i + 2a_2^i(x - x_{i-1}) + 3a_3^i(x - x_{i-1})^2,$$

и из условий $\left(H_3^i\right)'(x_{i-1}) = y'_{i-1}$, $\left(H_3^i\right)'(x_i) = y'_i$ получаем еще два уравнения:

$$a_1^i = y'_{i-1}, \quad (11.70)$$

$$a_1^i + 2a_2^i h_i + 3a_3^i h_i^2 = y'_i. \quad (11.71)$$

Равенства (11.68) и (11.70) уже определяют первые коэффициенты многочлена $H_3^i(x)$. Подставив их в оставшиеся равенства (11.69) и (11.71), легко находим третий и четвертый коэффициенты:

$$a_2^i = \frac{1}{h_i^2}(3y_i - 3y_{i-1} - 2h_i y'_{i-1} - h_i y'_i), \quad (11.72)$$

$$a_3^i = \frac{1}{h_i^3}(2y_{i-1} - 2y_i + h_i y'_{i-1} + h_i y'_i). \quad (11.73)$$

*) Здесь i — определяемый номером узла индекс, а не показатель степени.

Обратим внимание на то, что формула (11.66) определяет эрмитов сплайн заведомо нечетной степени. Для построения эрмитовых сплайнов четных степеней постановка задачи несколько видоизменяется.

Как и в случае рассматриваемых ранее (см. § 11.3) квадратичных интерполяционных сплайнов, будем строить эрмитов сплайн $S_{2m}(x)$ для заданной таблицей (11.65) функции $y = f(x)$, вводя дополнительные узлы z_i между данными узлами интерполяции x_{i-1} и x_i; для простоты будем считать, что z_i находится точно посередине между x_{i-1} и x_i. Тогда одно звено $G_{2m}^i(x)$ сплайна $S_{2m}(x)$ на элементарном промежутке $[x_{i-1}, x_i]$ составляется из двух многочленов:

$$G_{2m}^i(x) := \begin{cases} p_{2m}^i(x) := a_0^i + a_1^i(x - x_{i-1}) + \cdots + a_{2m}^i(x - x_{i-1})^{2m}, \\ \qquad \text{если } x \in [x_{i-1}, z_i], \\ q_{2m}^i(x) := b_0^i + b_1^i(x_i - x) + \cdots + b_{2m}^i(x_i - x)^{2m}, \\ \qquad \text{если } x \in [z_i, x_i]. \end{cases}$$

Для нахождения $2(2m+1)$ коэффициентов многочленов–полузвеньев $p_{2m}^i(x)$ и $q_{2m}^i(x)$ i-го звена $G_{2m}^i(x)$ сплайна $S_{2m}(x)$ (дефекта m) привлекаются $2m + 2$ интерполяционных условия

$$\left(G_{2m}^i\right)^{(j)}(x_{i-1}) = y_{i-1}^{(j)}, \quad \left(G_{2m}^i\right)^{(j)}(x_i) = y_i^{(j)} \quad (j = 0, 1, \ldots, m)$$

и $2m$ условий гладкой стыковки полузвеньев в точке z_i:

$$\left(p_{2m}^i\right)^{(j)}(z_i) = \left(q_{2m}^i\right)^{(j)}(z_i) \qquad (j = 0, 1, \ldots, 2m-1).$$

Несложно доказать, что такие коэффициенты могут быть однозначно найдены без привлечения информации с других промежутков, что означает возможность построения локальных сплайнов четной степени.

Построим **квадратичный эрмитов сплайн**, пользуясь теми же данными, которые использовались при построении кубического эрмитова сплайна.

Представляя i-е звено эрмитова сплайна $S_2(x)$ многочленами

$$p_2^i(x) = a_0^i + a_1^i(x - x_{i-1}) + a_2^i(x - x_{i-1})^2, \text{ если } x \in [x_{i-1}, z_i], \quad (11.74)$$

и

$$q_2^i(x) = b_0^i + b_1^i(x_i - x) + b_2^i(x_i - x)^2, \quad \text{если } x \in [z_i, x_i], \quad (11.75)$$

дифференцируем их:

$$\left(p_2^i\right)'(x) = a_1^i + 2a_2^i(x - x_{i-1}),$$

$$\left(q_2^i\right)'(x) = -b_1^i - 2b_2^i(x_i - x).$$

Далее на основе равенств

$$p_2^i(x_{i-1}) = y_{i-1}, \quad (p_2^i)'(x_{i-1}) = y'_{i-1},$$

$$q_2^i(x_i) = y_i, \quad (q_2^i)'(x_i) = y'_i,$$

$$p_2^i(z_i) = q_2^i(z_i), \quad (p_2^i)'(z_i) = (q_2^i)'(z_i),$$

полагая $z_i - x_{i-1} = x_i - z_i = \frac{1}{2}h_i$, устанавливаем связи между известными величинами y_{i-1}, y_i, y'_{i-1}, y'_i и неизвестными коэффициентами a_0^i, a_1^i, a_2^i, b_0^i, b_1^i, b_2^i i-го звена искомого сплайна:

$$\begin{cases} a_0^i = y_{i-1}, \\ a_1^i = y'_{i-1}, \\ b_0^i = y_i, \\ b_1^i = -y'_i, \\ a_0^i + \frac{1}{2}h_i a_1^i + \frac{1}{4}h_i^2 a_2^i = b_0^i + \frac{1}{2}h_i b_1^i + \frac{1}{4}h_i^2 b_2^i, \\ a_1^i + h_i a_2^i = -b_1^i - h_i b_2^i. \end{cases} \qquad (11.76)$$

Первые четыре уравнения полученной системы (11.76) представляют собой явные выражения первых четырех коэффициентов многочленов (11.74) и (11.75); подставив их в последние два уравнения, находим оставшиеся два коэффициента:

$$a_2^i = \frac{2}{h_i^2}(y_i - y_{i-1}) - \frac{1}{2h_i}(y'_i + 3y'_{i-1}), \qquad (11.77)$$

$$b_2^i = \frac{2}{h_i^2}(y_{i-1} - y_i) + \frac{1}{2h_i}(3y'_i + y'_{i-1}). \qquad (11.78)$$

Пример 11.3. Рассмотрим приближение кубическим и квадратичным эрмитовыми сплайнами функции $y = \sin \pi x$ на отрезке $[0, 1]$, если информация о ней представлена следующей таблицей:

x	y	y'
0	0	π
$\frac{1}{2}$	1	0
1	0	$-\pi$

Полагая $i = 1$ и $i = 2$, по формулам (11.68), (11.70), (11.72) и (11.73) находим коэффициенты кубического эрмитова сплайна, подстановка которых в (11.67) дает

$$S_3(x) = \begin{cases} H_3^1(x) = \pi x + (12 - 4\pi)x^2 + (4\pi - 16)x^3 & \text{при } x \in \left[0, \frac{1}{2}\right], \\ H_3^2(x) = 1 + (2\pi - 12)\left(x - \frac{1}{2}\right)^2 + (16 - 4\pi)\left(x - \frac{1}{2}\right)^3 & \text{при } x \in \left[\frac{1}{2}, 1\right]. \end{cases}$$

Аналогично, с помощью вычисления коэффициентов по формулам (11.76)–(11.78) с последующей их подстановкой в (11.74)–(11.75), приходим к квадратичному эрмитову сплайну

$$S_2(x) = \begin{cases} p_2^1(x) = \pi x + (8 - 3\pi)x^2 & \text{при } x \in \left[0, \frac{1}{4}\right], \\ q_2^1(x) = 1 + (\pi - 8)\left(\frac{1}{2} - x\right)^2 & \text{при } x \in \left[\frac{1}{4}, \frac{1}{2}\right], \\ p_2^2(x) = 1 + (\pi - 8)\left(x - \frac{1}{2}\right)^2 & \text{при } x \in \left[\frac{1}{2}, \frac{3}{4}\right], \\ q_2^2(x) = \pi(1 - x) + (8 - 3\pi)(1 - x)^2 & \text{при } x \in \left[\frac{1}{2}, \frac{3}{4}\right]. \end{cases}$$

Точность аппроксимации функции $\sin \pi x$ и ее производной $\pi \cos \pi x$ с помощью построенных эрмитовых сплайнов $S_3(x)$, $S_2(x)$ и их производных покажем в двух контрольных точках следующей таблицей:

x	$\sin \pi x$	$S_3(x)$	$S_2(x)$	$(\sin \pi x)'$	$S'_3(x)$	$S'_2(x)$
$\frac{1}{6}$	0.5	$\frac{2\pi+7}{27} \approx$ ≈ 0.492	$\frac{2\pi+7}{27} \approx$ ≈ 0.492	$\frac{\pi\sqrt{3}}{2} \approx$ ≈ 2.72	$\frac{8}{3} \approx$ ≈ 2.67	$\frac{8}{3} \approx$ ≈ 2.67
$\frac{3}{4}$	$\frac{\sqrt{2}}{2} \approx$ ≈ 0.707	$\frac{\pi+8}{16} \approx$ ≈ 0.696	$\frac{\pi+8}{16} \approx$ ≈ 0.696	$\frac{\pi\sqrt{3}}{2} \approx$ ≈ 2.72	$\frac{\pi}{4} - 3 \approx$ ≈ -2.21	$\frac{\pi}{2} - 4 \approx$ ≈ -2.43

Как видно из итоговой таблицы примера 11.3, эрмитова сплайн-интерполяция дает неплохие результаты. Это хорошо согласуется с имеющимися утверждениями о сходимости эрмитовых сплайнов $S_{2m+1}(x)$ и $S_{2m}(x)$ к интерполируемым ими функциям $f(x) \in C^m[a,b]$ с оценкой погрешности $O\left((\max h_i)^{2m+1}\right)$ и $O\left((\max h_i)^{2m}\right)$ соответственно (см., например, [91, 167]); для более гладких функций $f(x)$ подобные утверждения устанавливают близость соответствующих производных до m-й степени включительно.

УПРАЖНЕНИЯ

11.1. Выведите формулу осреднения (линейный фильтр) по пяти точкам. Примените ее для сглаживания данных примера 11.2 § 11.1.Результаты сглаживания отобразите графически; проведите визуальное сравнение с результатами осреднения по трем точкам (см. рис.11.3).

11.2. Постройте линейный сплайн $S_1(x)$, интерполирующий функцию $f(x)$, заданную таблицей

x	0	1	2	3
$f(x)$	0	1	1	3

Выразите $S_1(x)$ через линейные B-сплайны (см.(11.62)).

11.3. По данным упр. 11.2 постройте кубический и квадратичный интерполяционные сплайны дефекта 1, полагая $f''(0) = f''(3) = 0$.

11.4. Выведите совокупность формул для построения квадратичного интерполяционного сплайна при краевых условиях типа (11.56) и постоянном шаге $x_k - x_{k-1} = h$, в основу которой легло бы решение трехдиагональной системы относительно коэффициентов b_i звеньев сплайна (11.29) (см. замечание 11.4).

11.5. Постройте квадратичный сплайн дефекта 1, узлы которого $x_0, x_1, \ldots, x_n$ совпадали бы с узлами интерполируемой им функции $f(x)$ [23].

11.6. Докажите справедливость формул, фигурирующих в записи кубического B-сплайна $B_{3,k}(x)$.

11.7. Запишите представления и системы уравнений для вычисления коэффициентов этих представлений эрмитовых сплайнов пятой и четвертой степеней.

ГЛАВА 12 ‖ЧИСЛЕННОЕ ИНТЕГРИРОВАНИЕ

Выводятся и изучаются методы приближенного вычисления определенных интегралов, опирающиеся на суммирование с теми или иными коэффициентами значений подынтегральной функции. А именно, рассматриваются квадратурные формулы прямоугольников, трапеций, Симпсона, Чебышева, Гаусса. Особое внимание уделяется построению эффективных алгоритмов, приводящих с минимальными вычислительными затратами к значению интеграла наперед заданной точности. В их основе — принцип Рунге практического оценивания точности двойным счетом с разными шагами и простые соотношения между некоторыми квадратурными формулами. Затрагивается также проблема вычисления несобственных интегралов.

12.1. ЗАДАЧА ЧИСЛЕННОГО ИНТЕГРИРОВАНИЯ. КВАДРАТУРНЫЕ ФОРМУЛЫ ПРЯМОУГОЛЬНИКОВ

Пусть требуется найти значение I интеграла Римана $\int_a^b f(x)dx$ для некоторой заданной на отрезке $[a, b]$ функции $f(x)$. Хорошо известно, что для функций, допускающих на промежутке $[a, b]$ конечное число точек разрыва первого рода, такое значение существует, единственно и может быть формально получено по определению:

$$I = \lim_{n\to\infty} \sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1}), \qquad (12.1)$$

где $\{x_i\}_{i=0}^n$ — произвольная упорядоченная система точек отрезка $[a, b]$ такая, что

$$\max\{x_0 - a,\ x_i - x_{i-1},\ b - x_n\} \to 0 \quad \text{при} \quad n \to \infty,$$

а ξ_i — произвольная точка элементарного промежутка $[x_{i-1}, x_i]$.

В математическом анализе обосновывается аналитический способ нахождения значения I с помощью знаменитой формулы Ньютона–Лейбница

$$I = F(b) - F(a), \qquad (12.2)$$

где $F(x)$ — некоторая первообразная для данной функции $f(x)$.

К сожалению, применение этого весьма привлекательного подхода к вычислению I наталкивается на несколько серьезных препятствий. Самое главное из них — это несуществование первообразной среди элементарных функций для большинства элементарных функций $f(x)$; например, таким способом не удается вычислить

$$\int_a^b \frac{\sin x}{x}dx, \quad \int_a^b \frac{dx}{\ln x}, \quad \int_a^b e^{-x^2}dx \quad \text{и т.п.}^{*)}$$

Если первообразная $F(x)$ для заданной функции $f(x)$ все же найдена, то вычисление двух ее значений $F(a)$ и $F(b)$ может оказаться более трудоемким, чем вычисление существенно большего количества значений $f(x)$.

Поскольку в общем случае значения функций находятся лишь приближенно, использование точной формулы (12.2) приводит к приближенному результату, который может быть более эффективно получен с помощью какой-либо специальной приближенной формулы на основе значений подынтегральной функции $f(x)$. Такие специальные приближенные формулы для вычисления определенных интегралов называют ***квадратурными формулами*** (***механическими квадратурами***) или ***формулами численного интегрирования***. Первый из этих терминов можно связать с геометрическим смыслом определенного интеграла: вычисление $I := \int_a^b f(x)dx$ при $f(x) \ge 0$ равносильно построению квадрата, равновеликого криволинейной трапеции с основанием $[a, b]$ и «крышей» $f(x)$.

Простые квадратурные формулы можно вывести непосредственно из определения интеграла, т.е. из представления (12.1). Зафиксировав там некоторое $n \ge 1$, будем иметь

$$I \approx \sum_{i=1}^{n} f(\xi_i)(x_i - x_{i-1}). \qquad (12.3)$$

Это приближенное равенство назовем ***общей формулой прямоугольников*** (площадь криволинейной трапеции приближенно заменяется площадью ступенчатой фигуры, составленной из прямоугольников, основаниями которых служат отрезки $[x_{i-1}, x_i]$, а высотами — ординаты $f(\xi_i)$, рис. 12.1).

*) По значимости с этим препятствием к принятию формулы (12.2) может конкурировать разве что возможная дискретность задания подынтегральной функции $f(x)$, довольно характерная для многих реальных приложений определенного интеграла.

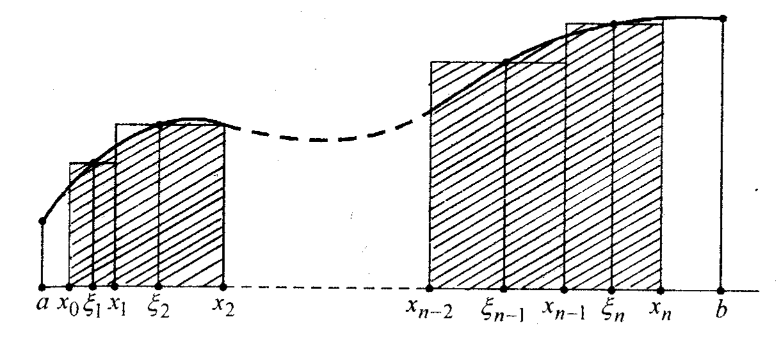


Рис. 12.1. *Геометрическая интерпретация общей формулы прямоугольников* (12.3)

Чтобы из общей формулы (12.3) получить конструктивное правило приближенного вычисления интеграла, воспользуемся свободой расположения точек x_i, разбивающих промежуток интегрирования $[a, b]$ на элементарные отрезки $[x_{i-1}, x_i]$, и свободой выбора точек ξ_i на этих отрезках.

Условимся в дальнейшем (до § 12.6) пользоваться равномерным разбиением отрезка $[a, b]$ на n частей точками x_i с шагом $h = \dfrac{b-a}{n}$, полагая

$$x_0 = a, \quad x_i = x_{i-1} + h \quad (i = 1, 2, \ldots, n-1), \quad x_n = b. \qquad (12.4)$$

При таком разбиении формула (12.3) приобретает вид

$$I \approx h\sum_{i=1}^{n} f(\xi_i), \quad \xi_i \in [x_{i-1}, x_i]. \qquad (12.5)$$

Теперь дело за фиксированием точек ξ_i на элементарных отрезках $[x_{i-1}, x_i]$. Рассмотрим три случая.

1) Положим $\xi_i = x_{i-1}$. Тогда из (12.5) получаем

$$I \approx I^{\Pi-} := h\sum_{i=1}^{n} f(x_{i-1}). \qquad (12.6)$$

2) Пусть в (12.5) $\xi_i = x_i$. Тогда имеем

$$I \approx I^{\Pi+} := h\sum_{i=1}^{n} f(x_i). \qquad (12.7)$$

Формулы (12.6) и (12.7) называются ***квадратурными формулами левых и правых прямоугольников*** соответственно. Совер-

шенно очевидно (рис. 12.2), что $I^{\Pi-}$ и $I^{\Pi+}$ дают двусторонние приближения к значению I интеграла от м о н о т о н н о й функции.

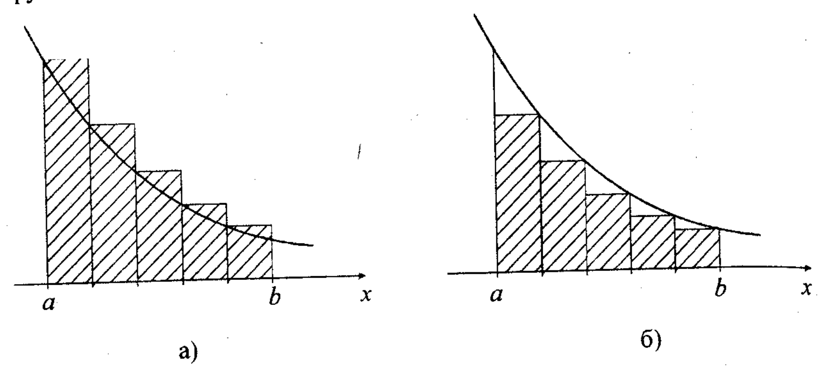


Рис. 12.2. Геометрическое оценивание интеграла от монотонной функции с помощью: а) $I^{\Pi-}$, б) $I^{\Pi+}$

Можно рассчитывать на бóльшую точность получения значения интеграла, если взять точку ξ_i посередине между точками x_{i-1} и x_i. Отсюда приходим к следующему случаю.

3) Фиксируем $\xi_i = \frac{1}{2}(x_{i-1} + x_i)$ $\left(= x_{i-1} + \frac{h}{2} = x_i - \frac{h}{2}\right)$. В результате имеем ***квадратурную формулу средних прямоугольников*** или, иначе, ***формулу средней точки***

$$I \approx I^{\Pi} := h\sum_{i=1}^{n} f\left(x_{i-1} + \frac{h}{2}\right) = h\sum_{i=1}^{n} f\left(x_i - \frac{h}{2}\right). \qquad (12.8)$$

Учитывая априорно бóльшую точность формулы (12.8) по сравнению с формулами (12.6) и (12.7) при том же объёме и характере вычислений, эту симметричную формулу будем впредь называть просто ***формулой прямоугольников***.

Полученные из определения интеграла квадратурные правила (12.6)–(12.8) не содержат в себе сведений, позволяющих сказать, каким нужно взять число n элементарных промежутков $[x_{i-1}, x_i]$ или, иначе, каким должен быть шаг h разбиения (12.4) отрезка интегрирования $[a, b]$, чтобы гарантированно найти значение I интеграла с наперед заданной точностью ε. Чтобы добыть эти сведения, положим вывод квадратурной формулы прямоугольников на другую основу, предъявив к подынтегральной функции $f(x)$ определенные требования.

Пусть функция $f(x)$ дважды непрерывно дифференцируема.

Рассмотрим сначала вычисление интеграла

$$I_0 := \int_{-\frac{h}{2}}^{\frac{h}{2}} f(x)dx \qquad (12.9)$$

при некотором достаточно малом $h > 0$.

По формуле Тейлора можно записать

$$f(x) = f(0) + f'(0)x + \frac{1}{2}f''(\xi)x^2,$$

где x — произвольная, а ξ — некоторая фиксированная точки интервала $\left(-\frac{h}{2}, \frac{h}{2}\right)$. Подставляя это выражение $f(x)$ в (12.9), имеем:

$$I_0 := \int_{-\frac{h}{2}}^{\frac{h}{2}} \left[f(0) + f'(0)x + \frac{1}{2}f''(\xi)x^2 \right]dx = hf(0) + \frac{f''(\xi_0)}{24}h^3, \quad (12.10)$$

где $\xi_0 \in \left(-\frac{h}{2}, \frac{h}{2}\right)$ — некоторая точка, вообще говоря, несовпадающая с точкой ξ (поскольку значение ξ изменяется с изменением h, к $f''(\xi)$ в интеграле $\int_{-\frac{h}{2}}^{\frac{h}{2}} f''(\xi)x^2\,dx$ нельзя относиться, как к постоянной, но, в силу неотрицательности x^2, можно применить теорему об интегральном среднем, согласно которой, если функция $f_2(x)$ сохраняет определенный знак на отрезке $[a, b]$, то в интервале (a, b) найдется точка c такая, что

$$\int_a^b f_1(x)f_2(x)\,dx = f_1(c)\int_a^b f_2(x)\,dx).$$

Слагаемое $hf(0)$ в выражении (12.10) интеграла I_0 очевидно, можно расценивать как приближение к I_0 по формуле прямоугольников в случае всего одного элементарного отрезка. Второе же слагаемое, т.е. $\frac{f''(\xi_0)}{24}h^3$ есть ***остаточный член (локальная погрешность) простейшей формулы прямоугольников*** $I_0 \approx hf(0)$.

Вернемся к нахождению значения I интеграла $\int_a^b f(x)\,dx$ при $f(x) \in C^2[a, b]$.

Выполним разбиение (12.4) отрезка $[a, b]$ на n частей с шагом $h = \frac{b-a}{n}$ и обозначим через $x_{i-\frac{1}{2}}$ середину i-го элементарного отрезка $[x_{i-1}, x_i]$. Тогда каждый отрезок $[x_{i-1}, x_i]$ представляется своим центром симметрии в виде $\left[x_{i-\frac{1}{2}} - \frac{h}{2}, x_{i-\frac{1}{2}} + \frac{h}{2}\right]$, и, следовательно, к каждому интегралу

$$I_i := \int_{x_{i-1}}^{x_i} f(x)\,dx = \int_{x_{i-\frac{1}{2}}-\frac{h}{2}}^{x_{i-\frac{1}{2}}+\frac{h}{2}} f(x)\,dx$$

применима формула (12.10) с $x_{i-\frac{1}{2}}$ в роли нуля, т.е.

$$I_i = hf(x_{i-\frac{1}{2}}) + \frac{f''(\xi_i)}{24}h^3,$$

где $i \in \{1, 2, \ldots, n\}$, $\xi_i \in (x_{i-1}, x_i)$.

Так как по свойству аддитивности, относящемуся к промежутку интегрирования,

$$\int_a^b f(x)\,dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)\,dx \quad , \qquad (12.11)$$

то

$$I = \sum_{i=1}^{n} I_i = \sum_{i=1}^{n} hf(x_{i-\frac{h}{2}}) + \sum_{i=1}^{n} \frac{f''(\xi_i)}{24}h^3 .$$

В первом слагаемом здесь узнаём выведенную ранее из других соображений формулу прямоугольников I^{Π} (сравните с (12.8)), второе же слагаемое характеризует остаточный член формулы прямоугольников

$$r^{\Pi}(h) := I - I^{\Pi} = \frac{h^3}{24}\sum_{i=1}^{n} f''(\xi_i).$$

Согласно обобщенной теореме о среднем значении непрерывной функции, существует точка $\xi_{\Pi} \in (a, b)$ такая, что

$$\sum_{i=1}^{n} f''(\xi_i) = \frac{1}{h}\sum_{i=1}^{n} hf''(\xi_i) = \frac{1}{h} f''(\xi_{\Pi})\sum_{i=1}^{n} h = \frac{b-a}{h} f''(\xi_{\Pi}).$$

В результате подстановки этого выражения в предыдущее равенство приходим к окончательному виду остаточного члена (***глобальной погрешности***) квадратурной формулы прямоугольников:

$$r^{\Pi}(h) = \frac{b-a}{24} f''(\xi_{\Pi})h^2 \ , \qquad \xi_{\Pi} \in (a, b). \qquad (12.12)$$

Как видно из формулы (12.12), при увеличении числа n элементарных отрезков, на которые разбивается промежуток интегрирования $[a, b]$, *ошибка численного интегрирования по формуле средней точки* (12.8) *убывает пропорционально квадрату шага* h. *Нетрудно убедиться, что погрешность численного интегрирования непрерывно дифференцируемой функции по формулам левых и правых прямоугольников* (12.6), (12.7) *убывает лишь по линейному закону.*

12.2. СЕМЕЙСТВО КВАДРАТУРНЫХ ФОРМУЛ НЬЮТОНА–КОТЕСА

Подстановка в интеграл $\int_a^b f(x)\,dx$ вместо функции $f(x)$ её интерполяционного многочлена Лагранжа той или иной степени n приводит к семейству квадратурных формул, называемых формулами Ньютона–Котеса*).

Как было показано в гл.1, функция $f(x) \in C^{n+1}[a, b]$ может быть единственным образом представлена в виде

$$f(x) = L_n(x) + R_n(x),$$

где $L_n(x) = \sum_{i=0}^{n} \frac{\Pi_{n+1}(x)y_i}{(x-x_i)\Pi'_{n+1}(x_i)}$ — интерполяционный многочлен Лагранжа, $R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\Pi_{n+1}(x)$ — остаточный член,

*) Ко́тес Роджер (1682–1716) — английский математик.

$y_i := f(x_i)$, $\Pi_{n+1}(x) := \prod_{i=0}^{n}(x - x_i)$, $\xi \in (a, b)$. Если система узлов интерполирования $\{x_i\}_{i=0}^{n}$ совпадает с точками разбиения (12.4) отрезка $[a, b]$ с шагом h, то замена переменной $x = x_0 + qh$ трансформирует многочлен Лагранжа к виду

$$L_n(x_0 + qh) = \sum_{i=0}^{n} \frac{(-1)^{n-i} y_i}{i!(n-i)!} \cdot \frac{q(q-1)\cdots(q-n)}{q-i}. \qquad (12.13)$$

Для того, чтобы использовать такое выражение $L_n(x)$ вместо $f(x)$ в $\int_a^b f(x)\,dx$, нужно изменить границы интегрирования (значению $x = a$ соответствует значение $q = 0$, а $x = b$ — значение $q = n$) и учесть, что $dx = hdq$. Таким образом, получаем

$$I \approx h \int_0^n \sum_{i=0}^{n} \frac{(-1)^{n-i} y_i}{i!(n-i)!} \cdot \frac{q(q-1)\cdots(q-n)}{q-i}\,dq.$$

Это равенство, переписанное в виде

$$I \approx (b-a) \sum_{i=0}^{n} H_i y_i, \qquad (12.14)$$

и есть ***квадратурная формула Ньютона–Котеса***, где

$$H_i := \frac{1}{n} \cdot \frac{(-1)^{n-i}}{i!(n-i)!} \int_0^n \frac{q(q-1)\ldots(q-n)}{q-i}\,dq \qquad (12.15)$$

— ***коэффициенты Котеса***.

На самом деле, формулы (12.14)–(12.15) определяют семейство квадратурных формул. Параметром этого семейства является число n — степень интерполяционного многочлена, которым заменяется подынтегральная функция.

Рассмотрим несколько простейших частных случаев, соответствующих небольшим значениям $n \in \mathbf{N}$. При этом конкретные формулы будем получать не на основе общих формул (12.14)–(12.15), а используя для этой цели вместо многочлена Лагранжа (12.13) эквивалентный ему (в силу единственности)

первый интерполяционный многочлен Ньютона (8.26):

$$P_n(x_0 + qh) = y_0 + q\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_0 + \ldots$$

$$\ldots + \frac{q(q-1)\ldots(q-n+1)}{n!}\Delta^n y_0. \qquad (12.16)$$

1) Пусть $n = 1$, т.е. имеется всего две точки x_0 и $x_1 = x_0 + h$, в которых известны значения функции ($y_0 = f(x_0)$ и $y_1 = f(x_1)$). Этим точкам соответствуют значения 0 и 1 переменной q. Следовательно,

$$\int_{x_0}^{x_1} f(x)\,dx \approx \int_0^1 (y_0 + q\Delta y_0) h\,dq = h\left[y_0 q + \frac{q^2}{2}\Delta y_0\right]_0^1 =$$

$$= h\left(y_0 + \frac{y_1 - y_0}{2}\right) = h\frac{y_0 + y_1}{2}. \qquad (12.17)$$

Получена ***простейшая квадратурная формула трапеций***, к которой легко прийти и из геометрических соображений (рис. 12.3).

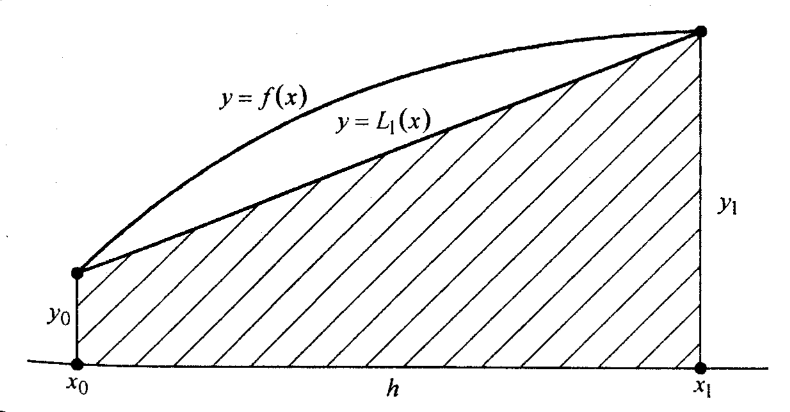


Рис. 12.3. Геометрическая интерпретация простейшей формулы трапеций

Остаточный член этой формулы найдем интегрированием остаточного члена $R_1(x)$ формулы линейной интерполяции, преобразованного к виду

$$R_1(x_0 + qh) = \frac{f''(\xi)}{2} h^2 q(q-1).$$

Имеем:

$$r_1 := \int_{x_0}^{x_1} f(x)\,dx - \frac{h}{2}(y_0 + y_1) = \frac{h^3}{2}\int_0^1 f''(\xi)\left(q^2 - q\right)dq =$$

$$= \frac{h^3}{2} f''(\xi_1)\int_0^1\left(q^2 - q\right)dq = -\frac{f''(\xi_1)}{12}h^3. \quad (12.18)$$

В найденном выражении остаточного члена $\xi_1 \in (x_0, x_1)$ — некоторая точка, отвечающая упоминавшейся в предыдущем параграфе интегральной теореме о среднем, которая здесь также применима, в силу неположительности функции $q^2 - q$ на отрезке $[0, 1]$.

2) Положим в (12.16) $n = 2$, т.е. проинтерполируем функцию $f(x)$ по трем точкам: x_0, $x_1 = x_0 + h$ и $x_2 = x_0 + 2h$. Тогда

$$\int_{x_0}^{x_2} f(x)\,dx \approx \int_0^2\left[y_0 + q\Delta y_0 + \frac{q(q-1)}{2}\Delta^2 y_0\right]h\,dq =$$

$$= h\left[2y_0 + 2(y_1 - y_0) + \frac{1}{3}(y_2 - 2y_1 + y_0)\right] = \frac{h}{3}(y_0 + 4y_1 + y_2). \quad (12.19)$$

Полученное приближенное равенство назовем ***простейшей формулой Симпсона****).

Поставим задачу найти ***остаточный член***

$$r_2 := \int_{x_0}^{x_2} f(x)\,dx - \frac{h}{3}(y_0 + 4y_1 + y_2) \quad (12.20)$$

простейшей формулы Симпсона (12.19). Так как функция $q(q-1)(q-2)$ меняет знак на промежутке $[0, 2]$, то здесь нельзя воспользоваться интегральной теоремой о среднем при интегрировании остаточного члена $R_2(x_0 + qh) = \frac{f'''(\xi)}{3!}h^3 q(q-1)(q-2)$ формулы квадратичной интерполяции, как это было в случае 1). Поэтому для выявления вида r_2 изберем искусственный путь.

Для удобства рассмотрим применение простейшей квадратурной формулы Симпсона (12.19) к интегралу с симметричными границами:

$$\int_{-h}^{h} f(x)\,dx = \frac{h}{3}[f(-h) + 4f(0) + f(h)] + r(h).$$

*) Си́мпсон Томас (1710–1761) — английский математик; прошел путь от ткача до профессора военной академии.

При любом $t \in [0, h]$ ее остаточный член есть

$$r(t) := \int_{-t}^{t} f(\tau)\,d\tau - \frac{t}{3}[f(-t) + 4f(0) + f(t)]. \quad (12.21)$$

Введем в рассмотрение функцию

$$v(t) := r(t) - \left(\frac{t}{h}\right)^5 r(h) \quad (12.22)$$

и изучим поведение $v(t)$ и нескольких ее первых производных, предполагая, что исходная функция $f(x)$ четырежды дифференцируема на $[-h, h]$.

Так как $\frac{d}{dt}\int_{-t}^{t} f(\tau)\,d\tau = f(t) - (-1)f(-t)$, то

$$v'(t) = f(t) + f(-t) - \frac{1}{3}[f(-t) + 4f(0) + f(t)] -$$

$$-\frac{t}{3}[f'(t) - f'(-t)] - \frac{5t^4}{h^5}r(h).$$

Дальнейшее дифференцирование дает:

$$v''(t) = \frac{1}{3}f'(t) - \frac{1}{3}f'(-t) - \frac{t}{3}[f''(t) + f''(-t)] - \frac{20t^3}{h^5}r(h),$$

$$v'''(t) = -\frac{t}{3}[f'''(t) - f'''(-t)] - \frac{60t^2}{h^5}r(h).$$

Последнее, благодаря формуле конечных приращений Лагранжа, примененной к разности третьих производных (в квадратных скобках), можно переписать в виде

$$v'''(t) = -\frac{2}{3}t^2\left[f^{IV}(\xi) + \frac{90}{h^5}r(h)\right], \quad (12.23)$$

где ξ — некоторая точка из интервала $(-t, t)$.

Теперь обратимся к анализу функции $v(t)$ и ее производных.

Как видно из (12.21), $r(0) = 0$; поэтому и $v(0) = 0$. Подстановка $t = h$ в (12.22) также приводит к нулевому значению $v(t)$. Следовательно, к функции $v(t)$ на отрезке $[0, h]$ применима теорема Ролля, согласно которой существует точка $t_1 \in (0, h)$ такая, что $v'(t_1) = 0$ (рис. 12.4). Непосредственной подстановкой значения $t = 0$ в выражение $v'(t)$ убеждаемся, что $v'(0) = 0$. Это означает, что теорема Ролля применима и к функции $v'(t)$ на отрезке

$[0,\ t_1]$, т.е.

$$\exists\, t_2 \in (0,\ t_1):\quad v''(t_2) = 0.$$

Так как $v''(0) = 0$, то по той же теореме

$$\exists\, \Theta \in (0,\ t_2):\quad v'''(\Theta) = 0.$$

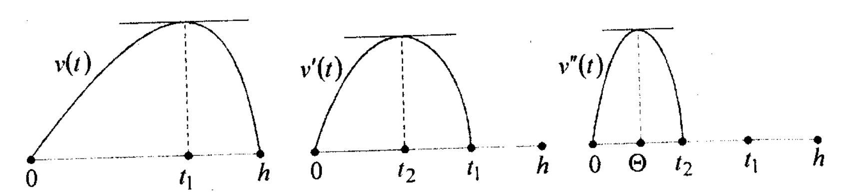


Рис. 12.4. *Геометрическая интерпретация трехкратного применения теоремы Ролля при выводе остаточного члена простейшей формулы Симпсона*

Таким образом, в соответствии с (12.23), при некотором $\Theta \in (0,\ t_2) \subset (0,\ h)$ справедливо равенство

$$-\frac{2}{3}\Theta^2\left[f^{IV}(\xi) + \frac{90}{h^5} r(h)\right] = 0,$$

из которого получаем выражение остаточного члена

$$r(h) = -\frac{h^5}{90} f^{IV}(\xi)\,,\quad \xi \in (-h,\ h).$$

Очевидно, эта формула может быть отнесена и к выражению r_2, определенному в (12.20), где промежуток интегрирования $[x_0,\ x_2]$ следует рассматривать как симметричный относительно точки x_1: $x \in [x_1 - h,\ x_1 + h]$; т.е. можно записать

$$r_2 = -\frac{h^5}{90} f^{IV}(\xi),\quad \xi \in (x_0,\ x_2). \tag{12.24}$$

3) Фиксируя степень $n = k$ интерполяционного многочлена (12.16) равной 3, 4, 5 и т.д., придем к частным формулам Ньютона–Котеса, подобным полученным выше простейшим формулам трапеций и Симпсона. Представим все эти частные случаи, включая уже рассмотренные, формулой вида

$$\int_{x_0}^{x_k} f(x)\,dx = B_k h \sum_{i=0}^{k} a_i^{(k)} f(x_i) + r_k(h), \tag{12.25}$$

где по-прежнему считается $x_i = x_0 + ih$, а коэффициенты B_k, $a_i^{(k)}$ и остаточные члены $r_k(h)$ задаются таблицей 12.1 (точка $\xi \in (x_0,\ x_k)$, разумеется, для каждого k своя).

Таблица 12.1

Параметры некоторых частных формул Ньютона–Котеса вида (12.25) [187]

k	B_k	$a_0^{(k)}$	$a_1^{(k)}$	$a_2^{(k)}$	$a_3^{(k)}$	$a_4^{(k)}$	$a_5^{(k)}$	…	$r_k(h)$
1	$\frac{1}{2}$	1	1						$-\frac{h^3}{12} f''(\xi)$
2	$\frac{1}{3}$	1	4	1					$-\frac{h^5}{90} f^{IV}(\xi)$
3	$\frac{3}{8}$	1	3	3	1				$-\frac{3h^5}{80} f^{IV}(\xi)$
4	$\frac{2}{45}$	7	32	12	32	7			$-\frac{8}{945} h^7 f^{VI}(\xi)$
5	$\frac{5}{288}$	19	75	50	50	75	19		$-\frac{275}{12096} h^7 f^{VI}(\xi)$
…	…	…	…	…	…	…	…	…	…

Замечание 12.1. В отличие от рассмотренных представлений формулы Ньютона–Котеса (12.14) или (12.25) непосредственно через значения подынтегральной функции, имеется другой ее вид, который можно назвать конечноразностным и который известен как ***формула Грегори****) [19, 187]:

$$\int_{x_0}^{x_0+kh} f(x)\,dx \approx h\Bigg[\left(\frac{1}{2}y_0 + y_1 + \ldots + y_{k-1} + \frac{1}{2}y_k\right) + \frac{1}{12}(\Delta y_0 - \Delta y_{k-1}) -$$

$$-\frac{1}{24}\left(\Delta^2 y_0 - \Delta^2 y_{k-2}\right) + \frac{19}{720}\left(\Delta^3 y_0 - \Delta^3 y_{k-3}\right) - \frac{3}{160}\left(\Delta^4 y_0 - \Delta^4 y_{k-4}\right) + \ldots\Bigg].$$

Очевидно, эта формула имеет такие же достоинства, как и конечноразностные интерполяционные формулы (см. § 8.12): слагаемые здесь расположены в порядке убывания значимости. При этом фиксирование $k = 1,\ 2,\ \ldots$ определяет соответственно формулы трапеций, Симпсона и др.

*) Грéгори Джеймс (1638–1675) — шотландский математик и астроном.

12.3. СОСТАВНЫЕ КВАДРАТУРНЫЕ ФОРМУЛЫ ТРАПЕЦИЙ И СИМПСОНА

Применение формул Ньютона–Котеса высоких порядков, т.е. при относительно больших значениях параметра $k \in \mathbf{N}$ в (12.25), может быть оправданным лишь при достаточно высокой гладкости подынтегральной функции $f(x)$. Более употребительными являются квадратурные правила, получающиеся путем дробления промежутка интегрирования на большое число мелких частей, интегрирование на каждой из которых производится с помощью однотипных простейших формул невысокого порядка. Выведем два таких правила — трапеций и Симпсона.

1. Прежде всего, заметим, что простейшая формула трапеций (12.17) с остаточным членом (12.18) применительно к интегрированию по отрезку $[x_{i-1}, x_i]$ может быть записана в виде точного равенства

$$\int_{x_{i-1}}^{x_i} f(x)\,dx = \frac{y_{i-1}+y_i}{2}h - \frac{f''(\xi_i)}{12}h^3, \qquad (12.26)$$

где ξ_i — некоторая, вообще говоря, неизвестная точка интервала (x_{i-1}, x_i), а y_i, как и прежде, есть укороченная запись значения $f(x_i)$.

Выполняя разбиение (12.4) исходного промежутка интегрирования $[a, b]$ на n частей с шагом $h = \frac{b-a}{n}$ и применяя к каждой из частей, на которые по свойству аддитивности расчленяется исходный интеграл, формулу (12.26), будем иметь

$$\int_a^b f(x)\,dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)\,dx = \frac{h}{2}\sum_{i=1}^{n}(y_{i-1}+y_i) - \frac{h^3}{12}\sum_{i=1}^{n} f''(\xi_i). \qquad (12.27)$$

Отсюда следует, что искомое значение интеграла можно приближенно найти по формуле

$$I \approx I^T := h\left(\frac{y_0+y_n}{2} + y_1 + y_2 + \ldots + y_{n-1}\right), \qquad (12.28)$$

которую в дальнейшем будем называть просто ***формулой трапеций***, а погрешность приближенного равенства (12.28) можно охарактеризовать остаточным членом r^T, полученным упрощением последнего слагаемого в (12.27). По обобщенной теореме о среднем значении функции на отрезке существует точка $\xi_T \in (a, b)$ такая, что

$$\sum_{i=1}^{n} h\, f''(\xi_i) = f''(\xi_T)(b-a)$$

(площадь ступенчатой фигуры, составленной из прямоугольников с основаниями h и высотами $f''(\xi_i)$, можно отождествить с площадью одного прямоугольника с основанием $\sum_{i=1}^{n} h = b-a$ и высотой $f''(\xi_T)$). Таким образом, ***остаточный член формулы трапеций*** (12.28) есть

$$r^T := I - I^T = -\frac{b-a}{12}h^2 f''(\xi_T), \quad \xi_T \in (a, b). \qquad (12.29)$$

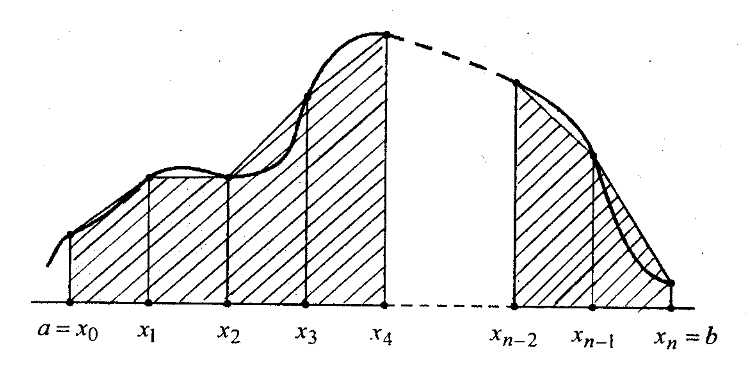


Рис. 12.5. Геометрическая интерпретация составной квадратурной формулы трапеций (12.28)

2. Аналогично равенству (12.26) на основе простейшей формулы Симпсона (12.19) и ее остаточного члена (12.24) запишем равенство

$$\int_{x_{2i-2}}^{x_{2i}} f(x)\,dx = \frac{h}{3}(y_{2i-2} + 4y_{2i-1} + y_{2i}) - \frac{h^5}{90} f^{(4)}(\xi_i), \qquad (12.30)$$
$$\xi_i \in (x_{2i-2}, x_{2i}).$$

Выполнив разбиение (12.4) так, чтобы число элементарных промежутков $n = 2m$ было четным, исходный интеграл представляем суммой m интегралов вида (12.30):

$$\int_a^b f(x)\,dx = \sum_{i=1}^{m} \int_{x_{2i-2}}^{x_{2i}} f(x)\,dx = \frac{h}{3}\sum_{i=1}^{m}(y_{2i-2} + 4y_{2i-1} + y_{2i}) - \frac{h^5}{90}\sum_{i=1}^{m} f^{(4)}(\xi_i).$$

Отсюда получается формула численного интегрирования

$$I \approx I^C := \frac{h}{3}(y_0 + y_{2m} + 4\sigma_1 + 2\sigma_2), \qquad (12.31)$$

где $\sigma_1 := y_1 + y_3 + ... + y_{2m-1}$, $\sigma_2 := y_2 + y_4 + ... + y_{2m-2}$, которая впредь будет называться ***формулой Симпсона***, и ее ***остаточный член***

$$r^C := I - I^C = -\frac{h^4}{180}\sum_{i=1}^{m} 2h\, f^{(4)}(\xi_i) = -\frac{b-a}{180}h^4 f^{(4)}(\xi_C) \qquad (12.32)$$

с некоторой точкой ξ_C из интервала (a, b).

Подводя итоги этого параграфа, можно сказать, что замена подынтегральной функции $f(x)$ на промежутке интегрирования $[a, b]$, разбитом на n частей с шагом h, кусочно-линейной функцией (рис. 12.5) приводит к его приближенному значению I^T (12.28) с ошибкой, убывающей при $h \to 0$ (согласно (12.29)) по квадратичному закону. Если же сделать кусочно-квадратичную интерполяцию подынтегральной функции по сдвоенным элементарным промежуткам $[x_0, x_2]$, $[x_2, x_4]$, ..., $[x_{2m-2}, x_{2m}]$, то ошибка получаемого таким образом приближенного равенства $I \approx I^C$ (12.31) в соответствии с (12.32) будет убывать пропорционально уже четвертой степени h. Это обстоятельство обусловило чрезвычайную популярность формулы Симпсона, ибо повышение порядка точности интерполяции всего на одну единицу повлекло повышение точности интегрирования на два порядка (при сохранении простоты расчетной формулы). Посмотрев на строку табл.12.1, соответствующую значению $k = 3$, легко понять, что применение кусочно-кубической интерполяции на основе формулы Ньютона–Котеса не даст принципиального улучшения качества соответствующей формулы численного интегрирования по сравнению с формулой Симпсона.

Пример 12.1. Оценим, какую точность можно гарантировать при вычислении интеграла

$$I = \int_0^1 e^{-x^2}\, dx$$

по формулам трапеций и Симпсона при разбиении промежутка интегрирования на восемь частей.

Выполнив последовательное дифференцирование функции $f(x) = e^{-x^2}$, находим

$$M_2 := \max_{x\in[a,\, b]}|f''(x)| = \max_{x\in[0,\, 1]}\left|2e^{-x^2}\left(2x^2 - 1\right)\right| = 2$$

и

$$M_4 := \max_{x\in[a,\, b]}\left|f^{(4)}(x)\right| = \max_{x\in[0,\, 1]}\left|4e^{-x^2}\left(4x^4 - 12x^2 + 3\right)\right| = 12.$$

Теперь, учитывая, что $b - a = 1$, $h = \dfrac{b-a}{n} = \dfrac{1}{8}$, в соответствии с формулами (12.29) и (12.32) получаем требуемые оценки:

$$\left|I - I^T\right| \le \frac{M_2(b-a)}{12}h^2 = \frac{2}{12}\left(\frac{1}{8}\right)^2 = \frac{1}{384} < 0.003$$

и

$$\left|I - I^C\right| \le \frac{M_4(b-a)}{180}h^4 = \frac{12}{180}\left(\frac{1}{8}\right)^4 = \frac{1}{61440} < 0.00002.$$

12.4. СООТНОШЕНИЯ МЕЖДУ ФОРМУЛАМИ ПРЯМОУГОЛЬНИКОВ, ТРАПЕЦИЙ И СИМПСОНА

Воспользоваться оценками квадратурных формул подобно тому, как это было сделано в примере 12.1, удается крайне редко. Поэтому вычисление интегралов с нужной точностью обычно производят посредством последовательного дробления шага (как правило, делением пополам) до выполнения некоторых критериев точности. Обоснование таких критериев и конкретные алгоритмы будут приведены в следующем параграфе. Здесь же изучим связи, которые возникают между приближенными значениями $I^П$, I^T и I^C интеграла I в процессе его вычисления по формулам прямоугольников, трапеций и Симпсона с шагом h и шагом $H = 2h$ (рис. 12.6).

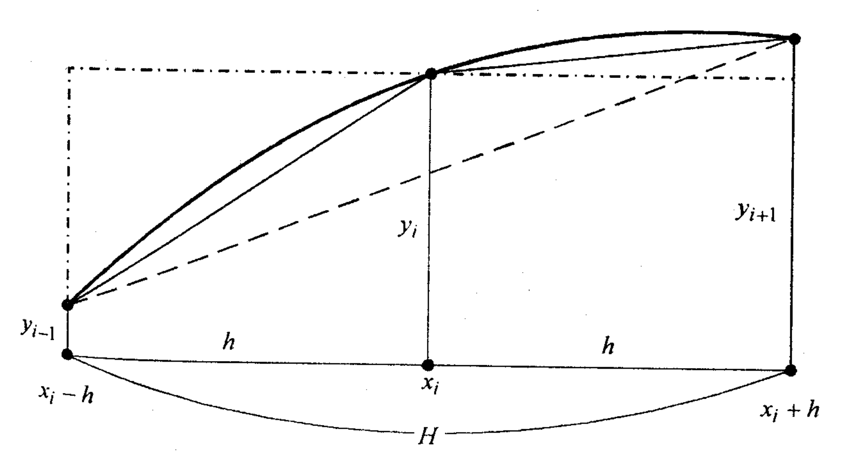


Рис. 12.6. К выводу соотношений между $I_i^П(H)$, $I_i^T(H)$, $I_i^T(h)$ и $I_i^C(h)$

Рассмотрим сначала приближения к интегралу

$$I_i := \int_{x_i-h}^{x_i+h} f(x)dx,$$

полученные по формуле прямоугольников с шагом H ($I_i^П(H)$), по формуле трапеций с шагами H и $h=\frac{H}{2}$ ($I_i^T(H)$ и $I_i^T(h)$), а также по формуле Симпсона с шагом h ($I_i^C(h)$), через значения $y_{i-1}=f(x_i-h)$, $y_i=f(x_i)$ и $y_{i+1}=f(x_i+h)$.

Имеем:

$$I_i^П(H)=Hy_i=2hy_i;$$

$$I_i^T(H)=\frac{y_{i-1}+y_{i+1}}{2}H=h(y_{i-1}+y_{i+1});$$

$$I_i^T(h)=\frac{y_{i-1}+y_i}{2}h+\frac{y_i+y_{i+1}}{2}h=\frac{h}{2}(y_{i-1}+2y_i+y_{i+1});$$

$$I_i^C(h)=\frac{h}{3}(y_{i-1}+4y_i+y_{i+1}).$$

Беря полусумму правых частей первых двух из этих равенств, получаем правую часть третьего. Следовательно,

$$I_i^T(h)=\frac{1}{2}\left(I_i^П(H)+I_i^T(H)\right). \qquad (12.33)$$

Если же взять среднее взвешенное этих же равенств с весами $\frac{2}{3}$ и $\frac{1}{3}$ соответственно, то придем к четвертому равенству, т.е.

$$I_i^C(h)=\frac{2}{3}I_i^П(H)+\frac{1}{3}I_i^T(H). \qquad (12.34)$$

Поскольку исходный интеграл есть $I=\sum_{i=1}^{m} I_i$, где $m=\frac{b-a}{H}$, то индекс i в соотношениях (12.33), (12.34) можно отбросить.

Итак, если $h=\frac{H}{2}$, то

$$I^T(h)=\frac{1}{2}\left(I^П(H)+I^T(H)\right), \qquad (12.35)$$

$$I^C(h)=\frac{2}{3}I^П(H)+\frac{1}{3}I^T(H). \qquad (12.36)$$

На базе этих связей можно строить эффективные алгоритмы вычисления интегралов. Суть в том, что, производя для уточнения значения интеграла дробление шага H пополам, мы должны вычислять новые значения подынтегральной функции в точках, расположенных посередине предыдущих элементарных отрезков. Эти новые значения используются для вычисления значения $I^П(H)$, все остальные значения функции передаются на этот этап с предыдущего уже в просуммированном виде, т.е. как $I^T(H)$. Так можно получать либо более точное, чем $I^T(H)$, значение $I^T(h)$, либо $I^C(h)$.

Заметим, кстати, что значение $I^C(h)$ можно вычислить также по формуле

$$I^C(h)=I^T(h)+R^T(h), \qquad (12.37)$$

где $R^T(h):=\frac{1}{3}\left(I^T(h)-I^T(H)\right)$ — так называемая ***поправка Ричардсона****). Действительно, возвращаясь к интегралу I_i, непосредственно (справа налево) проверяем справедливость i-го фрагмента формулы (12.37):

$$I_i^T(h)+R_i^T(h)=I_i^T(h)+\frac{1}{3}\left(I_i^T(h)-I_i^T(H)\right)=\frac{4}{3}I_i^T(h)-\frac{1}{3}I_i^T(H)=$$
$$=\frac{4}{3}\cdot\frac{h}{2}(y_{i-1}+2y_i+y_{i+1})-\frac{1}{3}\cdot h(y_{i-1}+y_{i+1})=\frac{h}{3}(y_{i-1}+4y_i+y_{i+1})=$$
$$=I_i^C(h);$$

из аддитивности интеграла теперь следует (12.37).

Таким образом, вычисление поправки Ричардсона позволяет уточнить приближенное значение $I^T(h)$ интеграла I, полученное по формуле трапеций. Другое ее назначение будет понятно из материала следующего параграфа.

12.5. ПРИНЦИП РУНГЕ ПРАКТИЧЕСКОГО ОЦЕНИВАНИЯ ПОГРЕШНОСТЕЙ. АЛГОРИТМ РОМБЕРГА

Пусть к приближенному вычислению значения I данного интеграла применяется некая квадратурная формула p-го порядка точности I^p из семейства составных формул Ньютона–

*) Ри́чардсон Льюис Фрай (1881–1953) — английский геофизик.

Котеса. При условии непрерывности p-ой производной подынтегральной функции это означает существование такой постоянной C, что

$$I = I^p(h) + Ch^p. \qquad (12.38)$$

При уменьшении вдвое шага h численного интегрирования по той же формуле p-го порядка можно записать такое же равенство, но с другой постоянной C_1:

$$I = I^p\left(\frac{h}{2}\right) + C_1\left(\frac{h}{2}\right)^p. \qquad (12.39)$$

Считая, что при малом h постоянные C и C_1 близки, из (12.38) и (12.39) имеем

$$I^p(h) + Ch^p = I^p\left(\frac{h}{2}\right) + C_1\left(\frac{h}{2}\right)^p \approx I^p\left(\frac{h}{2}\right) + C\left(\frac{h}{2}\right)^p$$

и, следовательно,

$$C \approx C_1 \approx \frac{I^p\left(\frac{h}{2}\right) - I^p(h)}{h^p - \left(\frac{h}{2}\right)^p}.$$

Подставив это значение C_1 в (12.39), приходим к выражению

$$I \approx I^p\left(\frac{h}{2}\right) + \frac{I^p\left(\frac{h}{2}\right) - I^p(h)}{2^p - 1}. \qquad (12.40)$$

К приближенному равенству (12.40) можно относиться двояко. С одной стороны, переписав его в виде

$$I - I^p\left(\frac{h}{2}\right) \approx \frac{I^p\left(\frac{h}{2}\right) - I^p(h)}{2^p - 1}, \qquad (12.41)$$

получаем возможность хотя бы грубо контролировать точность численного интегрирования на основе двойного счета (с шагом h и с шагом $\frac{h}{2}$). В этом и заключается широко применяемый ***принцип Рунге*** практического оценивания погрешностей. Его применение считается правомочным [44], если выполняется неравенство

$$\left| 2^p \cdot \frac{I^p\left(\frac{h}{2}\right) - I^p(h)}{I^p(h) - I^p(2h)} - 1 \right| < 0.1.$$

С другой стороны, второе слагаемое в формуле (12.40) позволяет уточнить «дешевым» способом приближенное значение $I^p\left(\frac{h}{2}\right)$ интеграла I. Заметим, что при $p = 2$ формула (12.38) есть формула трапеций

$$I = I^T(h) + Ch^2,$$

и выражение (12.41) в этом случае совпадает с определенной в предыдущем параграфе поправкой Ричардсона $R^T\left(\frac{h}{2}\right)$, а правая часть формулы (12.40) есть значение $I^C\left(\frac{h}{2}\right)$, соответствующее формуле Симпсона. Естественно назвать ***обобщенной поправкой Ричардсона*** получаемую с помощью двойного счета величину*)

$$R^p\left(\frac{h}{2}\right) := \frac{I^p\left(\frac{h}{2}\right) - I^p(h)}{2^p - 1}.$$

Итогом рассуждений предыдущего и настоящего параграфов может служить, например, следующий **алгоритм прямоугольников-трапеций** вычисления интеграла I с заданной точностью ε.

Шаг 1. Полагаем $n := 1$, $H := b - a$, $I^T(H) := \frac{H}{2}[f(a) + f(b)]$.

Шаг 2. Вычисляем $h = \frac{H}{2}$, $x_1 = a + h$, $x_i = x_{i-1} + H$

$(i = 2, \ldots, n)$, $y_i = f(x_i)$ $(i = 1, 2, \ldots, n)$, $I^{\Pi}(H) = H\sum_{i=1}^{n} y_i$.

Шаг 3. Вычисляем $I^T(h) = \frac{1}{2}\left[I^{\Pi}(H) + I^T(H)\right]$,

$$R^T(h) = \frac{1}{3}\left[I^T(h) - I^T(H)\right].$$

Шаг 4. Сравниваем $|R^T(h)|$ с ε. Если $|R^T(h)| > \varepsilon$, то полагаем $n := 2n$, $H := h$, $I^T(H) := I^T(h)$ и возвращаемся к шагу 2.

Шаг 5. Вычисляем $I^C(h) = I^T(h) + R^T(h)$ и принимаем $I :\approx I^C(h)$.

---

*) Не вызовет затруднений рассмотрение и более общего случая, когда шаг h заменяется на шаг $h_1 = \gamma h$, где $\gamma \in (0, 1)$.

Если есть основания считать, что подынтегральная функция обладает достаточно высокой степенью гладкости, то для вычисления интеграла можно применить алгоритм, предложенный шведским математиком Ромбергом. Базируется этот алгоритм на связях между составными квадратурными формулами Ньютона–Котеса четных порядков типа установленной выше связи (12.37) между формулами трапеций и Симпсона через поправку Ричардсона. Так, можно показать, что правая часть соответствующего (12.40) при $p = 4$ приближенного равенства

$$I \approx I^4\left(\frac{h}{2}\right) + \frac{I^4\left(\frac{h}{2}\right) - I^4(h)}{15}$$

(где $I^4(h) \equiv I^C(h)$, согласно предыдущему) совпадает со значением $I^6\left(\frac{h}{2}\right)$, которое могло бы быть подсчитано по составной формуле Ньютона–Котеса, отвечающей строке $k = 4$ в табл.12.1, и т.д.

Следовательно, заложив процесс деления шага пополам и выполнив первоначальный подсчет приближенного значения интеграла I по формуле трапеций, дальнейшие его уточнения можно производить рекуррентно путем прибавления обобщенных поправок Ричардсона к предыдущим приближениям к I.

Таким образом, **алгоритм Ромберга** определяется следующей совокупностью формул [13]:

$$h_0 := b - a,\ \ h_i = \frac{h_{i-1}}{2},\ \ I^{(0)}(h_0) := I^T(h_0),\ \ I^{(0)}(h_i) := I^T(h_i),$$

$$R^{(k-1)}(h_i) := \frac{1}{2^{2k} - 1}\left[I^{(k-1)}(h_i) - I^{(k-1)}(h_{i-1})\right],$$

$$I^{(k)}(h_i) = I^{(k-1)}(h_i) + R^{(k-1)}(h_i),$$

где i принимает значения $1, 2, 3, \ldots$, а k изменяется от 1 до i; счет заканчивается при выполнении условия $\left|R^{(k-1)}(h_i)\right| \le \varepsilon$, после чего полагают $I :\approx I^{(k)}(h_i)$.

Последовательность уточнений значения интеграла в алгоритме Ромберга удобно проследить по таблице 12.2 (см. нумерацию клеток таблицы).

Таблица 12.2

Схема последовательных уточнений интеграла алгоритмом Ромберга

i	h_i	$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	…
0	h_0	1 $I^{(0)}(h_0)$					
1	h_1	2 $I^{(0)}(h_1)$	3 $I^{(1)}(h_1)$				
2	h_2	4 $I^{(0)}(h_2)$	5 $I^{(1)}(h_2)$	6 $I^{(2)}(h_2)$			
3	h_3	7 $I^{(0)}(h_3)$	8 $I^{(1)}(h_3)$	9 $I^{(2)}(h_3)$	10 $I^{(3)}(h_3)$		
4	h_4	11 $I^{(0)}(h_4)$	12 $I^{(1)}(h_4)$	13 $I^{(2)}(h_4)$	14 $I^{(3)}(h_4)$	15 $I^{(4)}(h_4)$	
…	…	…	…	…	…	…	…

Наверное, не будет лишним обратить внимание на формирование столбца этой таблицы, соответствующего $k = 0$. Здесь должны находиться значения интеграла, получаемые по формуле трапеций с уменьшающимся в два раза шагом. Очевидно, к этому этапу целесообразно подключить часть приведенного выше алгоритма прямоугольников-трапеций, которая позволит экономично вычислять значения $I^{(0)}(h_i) = I^T(h_i)$ на основе соотношения (12.35).

12.6. КВАДРАТУРНЫЕ ФОРМУЛЫ ЧЕБЫШЕВА И ГАУССА

Общий вид линейной квадратурной формулы — это

$$\int_a^b f(x)\,dx \approx \sum_i A_i f(x_i), \tag{12.42}$$

где фиксированные аргументы x_i называют ***узлами***, а коэффициенты A_i — ***весами*** (***весовыми коэффициентами***) ***квадратурной формулы*** (определенный интеграл приближенно равен среднему взвешенному значений подынтегральной функции, вычисленных в определенных точках промежутка интегрирования).

Все рассмотренные выше квадратурные формулы характерны тем, что узлы в них брались равноотстоящими с шагом h, а веса находились в результате подмены подынтегральной функции $f(x)$ кусочно-постоянной в случае формул прямоугольников,

кусочно-линейной в случае формулы трапеций, кусочно-квадратичной в случае формулы Симпсона и т.д. Например, у составной формулы трапеций набор весов получился следующий:

$$\frac{h}{2},\ h,\ h,\ \ldots,\ h,\ \frac{h}{2},$$

а у составной формулы Симпсона —

$$\frac{h}{3},\ \frac{4h}{3},\ \frac{2h}{3},\ \frac{4h}{3},\ \frac{2h}{3},\ \ldots,\ \frac{4h}{3},\ \frac{h}{3}.$$

Далее мы откажемся от равномерного распределения узлов x_i на промежутке интегрирования $[a, b]$ и будем их н а х о д и т ь из тех или иных соображений. В таком случае целесообразно предварительно сделать линейную замену (см. § 9.2)

$$x = \frac{a+b}{2} + \frac{b-a}{2}t$$

и преобразовать исходный интеграл к интегралу со стандартным промежутком интегрирования $[-1, 1]$:

$$\int_a^b f(x)dx = \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{a+b}{2} + \frac{b-a}{2}t\right)dt. \qquad (12.43)$$

Равенство (12.43) позволяет без ограничения общности рассматривать теперь вычисление интеграла

$$\hat{I} \approx \int_{-1}^{1}\varphi(t)dt,$$

т.е. строить квадратурные формулы вида

$$\hat{I} \approx \sum_{i=1}^{n} A_i\varphi(t_i), \qquad (12.44)$$

от которых на основе (12.43) легко перейти к квадратурным формулам (12.42).

Сделаем еще два предварительных замечания.

Во-первых, обратим внимание на разную роль параметра n в квадратурных формулах предыдущих параграфов и в формуле (12.44): если ранее n означало число элементарных отрезков, на которые разбивался промежуток интегрирования, то здесь n — это число узлов.

Во-вторых, квадратурные формулы, основанные на равномерном разбиении отрезка $[a, b]$ на n частей с шагом h, сравнивали по точности в зависимости от степени h, содержащейся в выражениях их остаточных членов (см. (12.12), (12.29), (12.32) и табл.12.1). Но те же остаточные члены позволяют судить о точности иначе. Поскольку в выражение остаточного члена квадратурной формулы входит множителем производная подынтегральной функции определенного, например, k-го порядка, и поскольку k-я производная многочлена $(k-1)$-й степени равна нулю, можно сказать, что соответствующая квадратурная формула точна для многочленов степени $k-1$. В таком случае $k-1$ есть ***алгебраический порядок точности*** квадратурной формулы.

Потребуем, чтобы в формуле (12.44) значения функции $\varphi(t_i)$ суммировались с одинаковыми весами $A_i \equiv A$, а узлы t_i получающейся при этом формулы

$$\int_{-1}^{1}\varphi(t)dt \approx A\sum_{i=1}^{n}\varphi(t_i) \qquad (12.45)$$

располагались на отрезке $[-1, 1]$ так, чтобы формула (12.45) была точна для многочленов степени n. Отвечающая этим условиям формула называется ***квадратурной формулой Чебышева*** (она хороша тем, что минимизирует вероятную ошибку, если значения $\varphi(t_i)$ имеют нормально распределенную случайную ошибку [90]).

Считая, что равенство (12.45) должно быть непременно точным для постоянной функции, подставим в него $\varphi(t) \equiv 1$. Имеем

$$\int_{-1}^{1}1dt = A\sum_{i=1}^{n}1, \quad \text{откуда } A = \frac{2}{n}.$$

Следовательно, вместо (12.45) теперь можно использовать более конкретный шаблон квадратурной формулы Чебышева

$$\int_{-1}^{1}\varphi(t)dt \approx \frac{2}{n}\sum_{i=1}^{n}\varphi(t_i). \qquad (12.46)$$

Чтобы выписать совокупность условий на расположение узлов t_i на отрезке $[-1, 1]$, можно либо подставить в левую и правую части равенства (12.46) функцию-многочлен

$$\varphi(t) = a_0 + a_1t + a_2t^2 + \ldots + a_nt^n$$

и считать это равенство точным при любых a_i, либо поочередно подставлять в (12.46) представителей многочленов, коими являются степенные функции $t, t^2, \ldots, t^n$ (функция $t^0 = 1$ уже использована для выявления веса $A = \frac{2}{n}$). В любом случае

приходим к нелинейной системе

$$\begin{cases} t_1 + t_2 + \ldots + t_n = 0, \\ t_1^2 + t_2^2 + \ldots + t_n^2 = \dfrac{n}{3}, \\ t_1^3 + t_2^3 + \ldots + t_n^3 = 0, \\ t_1^4 + t_2^4 + \ldots + t_n^4 = \dfrac{n}{5}, \\ \ldots\ldots\ldots\ldots \\ t_1^n + t_2^n + \ldots + t_n^n = \dfrac{1+(-1)^n}{2} \cdot \dfrac{n}{n+1}. \end{cases} \qquad (12.47)$$

Доказано, что эта специфическая симметричная система n уравнений с n неизвестными определяет единственный набор узлов $\{t_i\}_{i=1}^n$ при $n = 1, 2, 3, \ldots, 7, 9$ (Чебышев) и не имеет действительных решений при $n = 8$ и $n \geq 10$ (Бернштейн*)). Эти узлы подсчитаны с высокой точностью и затабулированы (при нескольких первых значениях n см. их далее в табл.12.3).

Вернувшись к более общему, чем (12.45), виду квадратурной формулы (12.44), замечаем, что эта формула имеет $2n$ параметров: n узлов t_i и n весов A_i. Если считать, что мы свободны в выборе как узлов, так и весов, можно попытаться подобрать их такими, чтобы равенство

$$\int_{-1}^{1} \varphi(t)dt \approx \sum_{i=1}^{n} A_i \varphi(t_i) \qquad (12.48)$$

было точным для многочленов степени $2n-1$ или, что то же, для $2n$ степенных функций $\varphi(t) = 1, t, t^2, \ldots, t^{2n-1}$. В этом случае построение квадратурной формулы наивысшего алгебраического порядка точности вида (12.48), называемой ***квадратурной формулой Гаусса***, упирается в решение нелинейной системы

$$\begin{cases} \sum A_i = 2, \quad \sum A_i t_i = 0, \\ \sum A_i t_i^2 = \dfrac{2}{3}, \quad \sum A_i t_i^3 = 0, \\ \ldots\ldots\ldots\ldots \\ \sum A_i t_i^{2n-2} = \dfrac{2}{2n-1}, \quad \sum A_i t_i^{2n-1} = 0, \end{cases} \qquad (12.49)$$

---

*) Бернштейн Сергей Натанович (1880–1968) — российский математик, академик АН СССР, один из основателей конструктивной теории функций.

(где i всюду изменяется от 1 до n), получающейся аналогично системе (12.47). Решение системы (12.49) весьма затруднительно, но его несложно обойти, если знать конечный результат.

Оказывается, узлами t_i квадратурной формулы Гаусса (12.48) служат корни многочлена Лежандра $\chi_n(t)$ (см. § 10.4), которые, как известно, существуют при любом n, различны и принадлежат интервалу $(-1, 1)$, а веса A_i находятся интегрированием базисных многочленов Лагранжа $l_i(t)$ степени $n-1$ (см. § 8.2), а именно:

$$A_i = \int_{-1}^{1} \frac{(t-t_1)\ldots(t-t_{i-1})(t-t_{i+1})\ldots(t-t_n)}{(t_i-t_1)\ldots(t_i-t_{i-1})(t_i-t_{i+1})\ldots(t_i-t_n)} dt. \qquad (12.50)$$

Докажем, что при таких t_i и A_i формула (12.48) будет точна при подстановке в нее вместо $\varphi(t)$ любого многочлена $P_{2n-1}(t)$ степени $2n-1$.

Согласно теореме о делении многочлена с остатком, для пары многочленов $P_{2n-1}(t)$ и $\chi_n(t)$ однозначно найдется другая пара многочленов $U_{n-1}(t)$ и $V_{n-1}(t)$ таких, что

$$P_{2n-1}(t) = U_{n-1}(t)\chi_n(t) + V_{n-1}(t). \qquad (12.51)$$

В силу линейной независимости и попарной ортогональности (см. (3.24)) системы многочленов Лежандра $\{\chi_j\}_{j=0}^n$, найдется такой набор коэффициентов $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$, что

$$U_{n-1}(t) = \alpha_0\chi_0(t) + \alpha_1\chi_1(t) + \ldots + \alpha_{n-1}\chi_{n-1}(t) \quad \forall\, t \in [-1, 1]$$

и

$$\int_{-1}^{1} U_{n-1}(t)\chi_n(t)dt = \alpha_0 \int_{-1}^{1}\chi_0(t)\chi_n(t)dt + \alpha_1 \int_{-1}^{1}\chi_1(t)\chi_n(t)dt + \ldots$$

$$\ldots + \alpha_{n-1}\int_{-1}^{1}\chi_{n-1}(t)\chi_n(t)dt = 0 \quad \forall\, \alpha_0, \alpha_1, \ldots, \alpha_{n-1}.$$

Следовательно, интегрирование (12.51) приводит к равенству

$$\int_{-1}^{1} P_{2n-1}(t)dt = \int_{-1}^{1} V_{n-1}(t)dt,$$

являющемуся результатом преобразования левой части формулы (12.48) при $\varphi(t) = P_{2n-1}(t)$.

Рассмотрим теперь правую часть формулы (12.48) с A_i, определяемыми посредством (12.50), и с t_i такими, что $\chi_n(t_i)=0$. Имеем:

$$\sum_{i=1}^{n} A_i P_{2n-1}(t_i) = \sum_{i=1}^{n} A_i U_{n-1}(t_i)\chi_n(t_i) + \sum_{i=1}^{n} A_i V_{n-1}(t_i) =$$

$$= \sum_{i=1}^{n} \int_{-1}^{1} \frac{(t-t_1)\dots(t-t_{i-1})(t-t_{i+1})\dots(t-t_n)}{(t_i-t_1)\dots(t_i-t_{i-1})(t_i-t_{i+1})\dots(t_i-t_n)} dt \cdot V_{n-1}(t_i) =$$

$$= \int_{-1}^{1}\left[\sum_{i=1}^{n} \frac{(t-t_1)\dots(t-t_{i-1})(t-t_{i+1})\dots(t-t_n)}{(t_i-t_1)\dots(t_i-t_{i-1})(t_i-t_{i+1})\dots(t_i-t_n)} \cdot V_{n-1}(t_i)\right]dt = \int_{-1}^{1} V_{n-1}(t)dt.$$

Последнее равенство в цепочке проведенных преобразований объясняется тем, что выражение в квадратных скобках под знаком интеграла есть интерполяционный многочлен Лагранжа $(n-1)$-й степени, составленный по n значениям $V_{n-1}(t_i)$ $(i=1, 2, \dots, n)$ многочлена $(n-1)$-й степени $V_{n-1}(t)$, т.е. этот многочлен однозначно (в силу единственности интерполяционного многочлена) восстанавливается интерполированием.

Итак, имеем точное равенство

$$\int_{-1}^{1} P_{2n-1}(t)dt = \sum_{i=1}^{n} A_i P_{2n-1}(t_i)$$

при указанном фиксировании величин A_i и t_i, означающее, что эти величины могут служить весами и узлами квадратурной формулы Гаусса (12.48).

Запишем теперь общую формулу для квадратур Чебышева и Гаусса применительно к исходному интегралу по промежутку $[a, b]$ на основе преобразования (12.43):

$$\int_a^b f(x)dx \approx \frac{b-a}{2}\sum_{i=1}^{n} A_i f\left(\frac{a+b}{2}+\frac{b-a}{2}t_i\right). \qquad (12.52)$$

Значения узлов t_i и весов A_i, округленные до шести знаков после запятой, при $n = 2, 3, 4, 5$ приведены в табл.12.3. При необходимости узнать t_i, A_i с большей точностью или при $n>5$ следует обратиться к другим источникам (например, [19, 90, 99, 104]).

Таблица 12.3

Узлы и веса квадратурных формул Чебышева и Гаусса

n	i	Формула Чебышева		Формула Гаусса	
		t_i	$A_i \equiv A$	t_i	A_i
2	1; 2	∓0.577350	1	∓0.577350	1
3	1; 3	∓0.707107	$\frac{2}{3}$	∓0.774597	$5/9$
	2	0		0	$8/9$
4	1; 4	∓0.794654	$\frac{1}{2}$	∓0.861136	0.347855
	2; 3	∓0.187592		∓0.339981	0.652145
5	1; 5	∓0.832497	$\frac{2}{5}$	∓0.906180	0.236927
	2; 4	∓0.374541		∓0.538469	0.478629
	3	0		0	0.568889

Остаточный член квадратурной формулы Гаусса (12.52), как следовало ожидать, выражается через $(2n)$-ю производную функции $f(x)$:

$$r_n^{Г} := I - I^{Г} = \frac{(b-a)^{2n+1}(n!)^4}{(2n+1)[(2n)!]^3} f^{(2n)}(\xi_Г), \quad \xi_Г \in (a, b); \quad (12.53)$$

его вывод можно найти, например, в [19]. Там же выводятся и остаточные члены $r_n^{Ч} := I - I^{Ч}$ квадратур Чебышева вида (12.46) для всех конкретных значений n, при которых они существуют. Для $r_n^{Ч}$, как и для остаточных членов формул Ньютона–Котеса, характерно «перепрыгивание» через порядок входящих в их выражение производных:

$$r_2^{Ч} = \frac{1}{135}\varphi^{IV}(\xi_2), \quad r_3^{Ч} = \frac{1}{360}\varphi^{IV}(\xi_3),$$

$$r_4^{Ч} = \frac{2}{42525}\varphi^{VI}(\xi_4), \quad r_5^{Ч} = \frac{13}{544320}\varphi^{VI}(\xi_5)$$

и т.д., где ξ_n — некоторые фиксированные точки интервала $(-1, 1)$. Сравнение этих величин с соответствующими значениями $r_n^{Г}$, поставляемыми формулой (12.53) при $a=-1$, $b=1$, т.е. для

квадратурной формулы Гаусса (12.48), а именно, со значениями

$$r_2^Г = \frac{1}{135}\varphi^{IV}(\tilde{\xi}_2), \quad r_3^Г = \frac{1}{15750}\varphi^{VI}(\tilde{\xi}_3),$$
$$r_4^Г = \frac{2}{3472875}\varphi^{VIII}(\tilde{\xi}_4), \quad r_5^Г = \frac{13}{1237732650}\varphi^{X}(\tilde{\xi}_5),$$

говорит о существенно более быстром убывании ошибки формулы Гаусса при достаточной гладкости интегрируемой функции.

Проблема корректного использования формул Чебышева и Гаусса (12.52) — в трудности оценивания модулей производных высоких порядков (если они еще существуют!). Выход можно искать в построении составных формул подобно тому, как это делалось выше. К сожалению, здесь неприменим напрямую принцип Рунге, которым можно было бы руководствоваться при разбиении интеграла на части для получения его значения с наперед заданной точностью. Полезную роль на этом пути может сыграть изучение ***квадратурных формул Маркова****) [9, 19, 129]. Алгебраическая точность таких формул ниже, чем у соответствующих формул Гаусса, но зато какие-то узлы могут произвольно фиксироваться. Например, в число узлов «насильно» могут быть включены одна или обе границы интегрирования; в последнем случае такая формула называется ***квадратурой Лобатто*****) [9, 13].

Замечание 12.2. По тому, входят границы интегрирования в число узлов или нет, квадратурные формулы подразделяются на ***формулы замкнутого и открытого типов*** соответственно. Очевидно, формулы Чебышева и Гаусса относятся к формулам открытого типа, а формулы Ньютона–Котеса — замкнутого. Введение квадратурных формул Маркова позволяет «замкнуть» формулы Гаусса.

Замечание 12.3. Если на промежутке интегрирования $[a, b]$ доступны значения не только подынтегральной функции, но и некоторых ее производных, можно отказаться от требования несовпадения узлов интегрирования, и от вида квадратурной формулы (12.42) перейти к более общему виду формул численного интегрирования с кратными узлами [104, 133]. В основу построения конкретных квадратурных формул такого типа можно положить, например, подмену подынтегральной функции $f(x)$ ее интерполяционным многочленом Эрмита $H_n(x)$ (см. § 8.8). Добавление к значениям функции только первых производных в точках a и b приводит к простейшей ***формуле Эйлера*** (Эйлера–Маклорена) [78]

$$I \approx \frac{b-a}{2}[f(a)+f(b)] + \frac{(b-a)^2}{12}[f'(a)-f'(b)],$$

на базе которой при равномерном разбиении (12.4) получается ***составная формула Эйлера***

$$I \approx I^Э := h\left(\frac{y_0+y_n}{2} + y_1 + y_2 + \ldots + y_{n-1}\right) + \frac{h^2}{12}(f'(a)-f'(b)). \qquad (12.54)$$

Последняя незначительно усложняет составную формулу трапеций (12.28), но существенно повышает ее точность, о чем свидетельствует величина ошибки $r^Э := I - I^Э \approx \frac{h^4}{720}\int_a^b f^{IV}(x)dx$ в сравнении с величиной $r^Т$ в (12.29). Интересно отметить, что формула (12.54) может быть выведена из формулы Грегори (см. замечание 12.1), если запись последней оборвать на втором слагаемом $\frac{1}{12}(\Delta y_0 - \Delta y_{k-1})$, зафиксировать $k = n$ и воспользоваться обсуждавшейся в первой главе связью между конечными разностями и производными: $\Delta y_0 \approx hf'(a)$, $\Delta y_{n-1} \approx hf'(b)$.

*) Марков Андрей Андреевич (1856–1922) — широко известный русский математик, академик Петербургской академии наук.

**) Лобатто Рехюэл (1797–1866) — нидерландский математик.

12.7. ФОРМУЛЫ ГАУССА–КРИСТОФФЕЛЯ

Квадратурной формулой Гаусса–Кристоффеля*) [71, 78] (или формулой ***типа Гаусса*** [129], или просто ***Гаусса*** [13, 19, 158]) называют формулу наивысшего алгебраического порядка точности вида

$$\int_a^b p(x)f(x)dx \approx \sum_{i=1}^{n} A_i f(x_i), \qquad (12.55)$$

где границы интегрирования a и b могут быть как конечными, так и бесконечными, а ***весовая функция*** $p(x)$ должна удовлетворять нескольким условиям. А именно, функция $p(x)$ должна быть непрерывна и положительна на интервале (a, b); при $x = a$ и $x = b$ она может обращаться в нуль или в бесконечность, при этом должен существовать $\int_a^b p(x)dx$. Очевидно, на конечном промежутке $[a, b]$ всем этим требованиям удовлетворяет функ-

*) Кристоффель Элвин Бруно (1829–1900) — немецкий математик.

ция $p(x) \equiv 1$, с которой квадратурная формула Гаусса вида (12.42) (или вида (12.48) при $a = -1$, $b = 1$) является частным случаем формулы Гаусса–Кристоффеля (12.55).

Обобщим знания о квадратурной формуле Гаусса, полученные в предыдущем параграфе, на формулу (12.55). При этом, естественно, будем пользоваться более широким понятием ортогональности с весом, согласно которому множество многочленов $\{Q_0(x), Q_1(x), \ldots, Q_n(x), \ldots\}$ образует систему ортогональных с весовой функцией $p(x)$ многочленов на (a, b), если

$$\int_a^b p(x)Q_k(x)Q_j(x)dx = \begin{cases} 0 & \text{при } k \neq j, \\ \text{const} \neq 0 & \text{при } k = j. \end{cases} \qquad (12.56)$$

Как уже упоминалось в § 10.4, многочлен n-й степени $Q_n(x)$ из такой системы имеет на промежутке ортогональности (a, b) ровно n вещественных корней (независимо от того, конечен этот промежуток или бесконечен). Они и принимаются за узлы квадратурной формулы Гаусса–Кристоффеля.

Теорема 12.1. *Квадратурная формула* (12.55) *точна для произвольного многочлена степени* $2n-1$, *если ее узлами* x_i *служат корни многочлена* $Q_n(x)$ *из семейства многочленов, ортогональных на промежутке интегрирования* (a, b) *с весом* $p(x)$, *а весовыми коэффициентами — числа*

$$A_i = \int_a^b \frac{(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)}\, p(x)dx. \qquad (12.57)$$

Доказательство проведем схематично, поскольку оно повторяет рассуждения, которые были положены в обоснование аналогичного результата для формулы Гаусса (12.48).

Возьмем произвольный многочлен степени $2n-1$ и представим его в виде

$$P_{2n-1}(x) = U_{n-1}(x)Q_n(x) + V_{n-1}(x) \qquad (12.58)$$

(сравните с (12.51)). Разложим $U_{n-1}(x)$ по базису $Q_0(x), Q_1(x), \ldots$ $\ldots, Q_{n-1}(x)$ с коэффициентами $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$ и подставим $P_{2n-1}(x)$ в качестве $f(x)$ в левую часть формулы (12.55). Тогда, в силу (12.56), будем иметь

$$\int_a^b p(x)P_{2n-1}(x)dx = \alpha_0 \int_a^b p(x)Q_0(x)Q_n(x)dx +$$
$$+ \alpha_1 \int_a^b p(x)Q_1(x)Q_n(x)dx + \ldots + \alpha_{n-1} \int_a^b p(x)Q_{n-1}(x)Q_n(x)dx +$$
$$+ \int_a^b p(x)V_{n-1}(x)dx = \int_a^b p(x)V_{n-1}(x)dx.$$

Подстановка в правую часть (12.55) функции $f(x) = P_{2n-1}(x)$ в виде (12.58) дает тот же результат:

$$\sum_{i=1}^n A_i P_{2n-1}(x_i) = \sum_{i=1}^n A_i U_{n-1}(x_i)Q_n(x_i) + \sum_{i=1}^n A_i V_{n-1}(x_i) = 0 +$$
$$+ \int_a^b \left[\sum_{i=1}^n \frac{(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)} \cdot V_{n-1}(x_i) \right] p(x)dx =$$
$$= \int_a^b p(x)V_{n-1}(x)dx,$$

так как по условию $Q_n(x_i) = 0$, а

$$\sum_{i=1}^n \frac{(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\ldots(x-x_n)}{(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\ldots(x_i-x_n)} \cdot V_{n-1}(x_i) = V_{n-1}(x),$$

согласно теории интерполирования.

Как видим, для алгебраической функции $P_{2n-1}(x)$ равенство (12.55) является точным. Теорема доказана.

Условия на узлы x_i и веса A_i фигурируют в теореме 12.1 в качестве достаточных условий для того, чтобы формула (12.55) являлась квадратурной формулой наивысшего алгебраического порядка точности. Доказано [99 и др.], что они являются и необходимыми.

Замечание 12.4. При любом расположении узлов x_i на $[a, b]$ формулу вида (12.55) называют ***интерполяционной квадратурной формулой*** [129] или ***квадратурной формулой интерполяционного типа*** [9, 158], если ее весовые коэффициенты задаются равенством (12.57). Таковыми являются не только рассматриваемые здесь формулы Гаусса–Кристоффеля, но и изученные ранее простейшие формулы прямоугольников, трапеций, Симпсона, а также формулы Чебышева. Интерполяцион-

ными квадратурами не будут формулы численного интегрирования, опирающиеся, например, на аппроксимацию подынтегральной функции многочленами наилучших среднеквадратических приближений (см. гл. 10), а не на интерполяцию.

Несколько конкретных квадратурных формул Гаусса–Кристоффеля можно получить на базе приведенных в § 10.4 классических ортогональных многочленов.

Положив в формуле (12.55) $a=-1,\ b=1,\ p(x)=\frac{1}{\sqrt{1-x^2}}$ и взяв в качестве узлов x_i корни многочлена Чебышева $T_n(x)$ (см. (2.3)), придем к формуле

$$\int_{-1}^{1}\frac{f(x)}{\sqrt{1-x^2}}dx \approx \sum_{i=1}^{n} A_i f(x_i), \tag{12.59}$$

которую называют ***квадратурной формулой Эрмита*** [19, 169][*]. Формула (12.59) имеет алгебраический порядок точности $2n-1$, когда ее коэффициенты A_i, согласно теореме 12.1, вычисляются по формуле 12.57. Учитывая краткую запись интерполяционного многочлена Лагранжа в виде (8.6а), формулу для вычисления весовых коэффициентов в данном случае можно представить так:

$$A_i = \int_{-1}^{1}\frac{T_n(x)}{(x-x_i)T_n'(x_i)}\cdot\frac{dx}{\sqrt{1-x^2}}. \tag{12.60}$$

Знание вида и свойств многочлена Чебышева позволяет найти значения интеграла в выражении (12.60) при $i=1,2,\dots,n$ и установить, что все коэффициенты в формуле (12.59) равны между собой [19, 99, 129]:

$$A_i=\frac{\pi}{n}\quad \forall\ i\in\{1,2,\dots,n\}.$$

Таким образом, квадратурная формула Эрмита есть

$$\int_{-1}^{1}\frac{f(x)dx}{\sqrt{1-x^2}}dx \approx \frac{\pi}{n}\sum_{i=1}^{n} f(x_i), \tag{12.61}$$

[*] В [90] так называется другая формула, см. далее (12.64).

где $x_i=\cos\frac{2i-1}{2n}\pi$ — корни многочлена $T_n(x)$. Известна она и как ***формула Мелера*** [111, 129].

Промежуточное положение между общей формулой Гаусса–Кристоффеля (12.55) и ее частными представителями, каковыми являются формула Гаусса и формула Эрмита, занимает квадратурная формула вида

$$\int_{-1}^{1}(1-x)^{\alpha}(1+x)^{\beta} f(x)dx \approx \sum_{i=1}^{n} A_i f(x_i)\quad (\alpha,\beta>-1). \tag{12.62}$$

На самом деле (12.62) — это двупараметрическое семейство квадратурных формул, имеющих наивысший алгебраический порядок точности, когда за узлы x_i принимаются корни ортогональных с фигурирующей под знаком интеграла весовой функцией многочленов Якоби (упоминавшихся в § 10.4), а коэффициенты A_i вычисляются в соответствии с формулой (12.57). При некоторых конкретных значениях параметров α и β узлы x_i и веса A_i можно найти затабулированными в справочной литературе (например, [104], где вместо $[-1,1]$ за основу принят промежуток $[0,1]$; там же содержатся числовые данные и для других формул численного интегрирования).

В случае произвольного з а м к н у т о г о промежутка интегрирования $[a,b]$ для того, чтобы воспользоваться квадратурными формулами типа формулы (12.62), достаточно свести его, как это делалось ранее в § 12.6, к стандартному промежутку $[-1,1]$.

Если одна или обе границы интегрирования бесконечны, тогда на основе ортогональных многочленов Лагерра и Эрмита приходим соответственно к формулам

$$\int_{0}^{\infty}e^{-x}f(x)dx \approx \sum_{i=1}^{n} A_i f(x_i) \tag{12.63}$$

и

$$\int_{-\infty}^{+\infty}e^{-x^2}f(x)dx \approx \sum_{i=1}^{n} A_i f(x_i), \tag{12.64}$$

которые иногда называют также соответственно ***квадратурными***

формулами Лагерра и Эрмита [90][*]. За узлы в них принимаются корни многочленов Лагерра $L_n(x)$ в (12.63) и Эрмита $H_n(x)$ в (12.64) (см. § 10.4), а весовые коэффициенты A_i подсчитываются по формулам [99]:

$$A_i = \int_0^{\infty} \frac{L_n(x)}{(x - x_i)L_n'(x_i)} e^{-x} dx = \frac{(n!)^2}{x_i [L_n'(x_i)]^2} = \left[\frac{(n-1)!}{nL_{n-1}(x_i)}\right]^2 x_i$$

для квадратуры (12.63) и

$$A_i = \int_{-\infty}^{+\infty} \frac{H_n(x)}{(x - x_i)H_n'(x_i)} e^{-x^2} dx = \frac{2^{n+1} n! \sqrt{\pi}}{[H_n'(x_i)]^2} = \frac{2^{n-1}(n-1)!\sqrt{\pi}}{nH_{n-1}^2(x_i)}$$

для (12.64) (точки x_i в этих формулах предполагаются уже фиксированными указанным образом)[**].

Числовые данные об узлах и весах квадратур Лагерра и Эрмита для значений n от 2 до 5 представлены в таблице 12.4 [90, 104].

Остаточные члены формул (12.63) и (12.64) имеют соответственно вид [99, 104]

$$R_n^{Л} = \frac{(n!)^2}{(2n)!} f^{(2n)}(\xi_{Л}) \quad \text{и} \quad R_n^{Э} = \frac{n!\sqrt{\pi}}{2^n (2n)!} f^{(2n)}(\xi_{Э}),$$

где $\xi_{Л} \in (0, +\infty)$, $\quad \xi_{Э} \in (-\infty, +\infty)$.

---

*) Называют их также ***квадратурными формулами с весом Чебышева–Лагерра*** и ***Чебышева–Эрмита*** соответственно [89]. Имеется и более общая ***формула Чебышева–Лагерра*** [104]

$$\int_0^{\infty} x^{\alpha} e^{-x} f(x) dx \approx \sum_{i=1}^{n} A_i f(x_i),$$

поглощающая (12.63) при $\alpha = 0$.

**) Чтобы в представленных здесь интегральных выражениях A_i узнать обоснованную в теореме 12.1 общую формулу (12.57), достаточно вспомнить краткую запись интерполяционного многочлена Лагранжа в виде (8.6а). Использование многочлена Лагерра $L_n(x)$ в первом и Эрмита $H_n(x)$ во втором интегральных выражениях подчеркивает, что за узлы $x_1, \ldots, x_n$ принимаются корни именно этих многочленов.

Таблица 12.4

Узлы и веса квадратурных формул Лагерра (12.63) *и Эрмита* (12.64)

n	i	Формула Лагерра		Формула Эрмита	
		x_i	A_i	x_i	A_i
2	1	0.585786	0.853553	–0.707107	0.886227
	2	3.414214	0.146447	0.707107	0.886227
3	1	0.415775	0.711093	–1.224745	0.295409
	2	2.294280	0.278518	0	1.181636
	3	6.289945	0.0103893	1.224745	0.295409
4	1	0.322548	0.603154	–1.650680	0.0813128
	2	1.745761	0.357419	–0.524648	0.804914
	3	4.536620	0.0388879	0.524648	0.804914
	4	9.395071	0.000539295	1.650680	0.0813128
5	1	0.263560	0.521756	–2.020183	0.0199532
	2	1.413403	0.398667	–0.958572	0.393619
	3	3.596426	0.0759424	0	0.945309
	4	7.085810	0.00361176	0.958572	0.393619
	5	12.640801	0.00002337	2.020183	0.0199532

12.8. ПРИЕМЫ ПРИБЛИЖЕННОГО ВЫЧИСЛЕНИЯ НЕСОБСТВЕННЫХ ИНТЕГРАЛОВ

В этом параграфе рассмотрим вычисление интегралов вида

А)[*] $\displaystyle\int_0^{\infty} f(x)dx$,

Б) $\displaystyle\int_{-\infty}^{+\infty} f(x)dx$

---

*) Более общий случай интегрирования на промежутке $[a; +\infty)$ сводится к случаю А) или линейной заменой переменной, или представлением интеграла $\int_a^{\infty}$ в виде двух интегралов: $\int_a^0$ и $\int_0^{\infty}$, один из которых — определенный.

или

В) $\int\limits_a^b f(x)dx$

при наличии бесконечных разрывов у функции $f(x)$ в последнем интеграле либо в точке $x=a$, либо в точке $x=b$, либо, в общем случае, в некоторой точке $x=c\in[a, b]$. Вычисляемые несобственные интегралы, разумеется, предполагаются сходящимися.

Одним из источников получения численных значений несобственных интегралов А–В являются рассмотренные в предыдущем параграфе квадратурные формулы из класса формул Гаусса–Кристоффеля. Для их применения нужно выделить под интегралом подходящую весовую функцию и воспользоваться соответствующей квадратурой.

Так, в случае А можно записать

$$\int\limits_0^{\infty} f(x)dx = \int\limits_0^{\infty} e^{-x}e^{x}f(x)dx \approx \sum_{i=1}^{n} A_i e^{x_i} f(x_i), \qquad (12.65)$$

т.е. применить квадратурную формулу Лагерра (12.63).

Аналогично, для интеграла Б имеем

$$\int\limits_{-\infty}^{+\infty} f(x)dx = \int\limits_{-\infty}^{+\infty} e^{-x^2}e^{x^2}f(x)dx \approx \sum_{i=1}^{n} A_i e^{x_i^2} f(x_i),$$

где A_i и x_i должны соответствовать формуле Эрмита (12.64).

Пример 12.2. Точное значение несобственного интеграла $\int\limits_0^{\infty} \ln \operatorname{th}\frac{x}{2}\,dx$ равно $-\frac{\pi^2}{4}\approx -2.4674$ [59].

Применение к этому интегралу формулы (12.65), т.е. приближенного равенства

$$\int\limits_0^{\infty} \ln th\frac{x}{2}dx \approx \sum_{i=1}^{n} A_i\varphi(x_i),$$

где $\varphi(x):=e^{x}\ln \operatorname{th}\frac{x}{2}=\left(\ln\frac{1-e^{-x}}{1+e^{-x}}\right)\Big/e^{-x}$, с A_i и x_i из табл.12.4, дает:

при $n=2$

$A_1\varphi(x_1)+A_2\varphi(x_2)\approx 0.8536\cdot(-2.2562)+0.1464\cdot(-2.0007)\approx -2.219$;

при $n=4$

$$\sum_{i=1}^{4} A_i\varphi(x_i)\approx -(0.6032\cdot 2.5311+0.3574\cdot 2.0207+0.03889\cdot 2.0001+$$
$$+0.0005393\cdot 2.0004)\approx -2.328.$$

В промежуточных вычислениях здесь обращает на себя внимание незначительное изменение от узла к узлу абсолютных значений функции $\varphi(x)$; быстрое убывание модуля подынтегральной функции $f(x)$ принимает на себя весовая функция e^{-x}, что находит отражение в убывании весовых коэффициентов A_i с ростом i. Как отмечается в [104], убывание или ограниченность модуля функции $e^{x}f(x)$ в (12.65) ($\varphi(x)$ в данном примере) чреваты медленной сходимостью квадратурного процесса Лагерра, так как эта функция должна аппроксимироваться многочленами, а у них модули растут при $x\to\infty$. Отсюда — целесообразность введения множителя x^{α} (см. сноску на с. 500), который за счет подбора параметра α позволит регулировать поведение интегрируемой с весом $x^{\alpha}e^{-x}$ функции.

Так, в данном примере при $n=4$, $\alpha=-\frac{1}{2}$, т.е. с весом Чебышева-Лагерра $e^{-x}/\sqrt{x}$, нужно просуммировать значения функции

$$\psi(x):=\sqrt{x}\,e^{x}\ln \operatorname{th}\frac{x}{2}=\frac{\sqrt{x}}{e^{-x}}\ln\frac{1-e^{-x}}{1+e^{-x}}$$

в узлах $x_1=0.145304$, $x_2=1.339097$, $x_3=3.926964$ и $x_4=8.588636$ с весовыми коэффициентами $A_1=1.322294$, $A_2=0.415605$, $A_3=0.0341560$ и $A_4=0.000399208$ соответственно [104]. Вычислив значения $\psi(x_1)\approx -1.1566$, $\psi(x_2)\approx -1.7696$, $\psi(x_3)\approx -3.9638$, $\psi(x_4)\approx -5.8581$ (возрастающие по модулю!), получаем приближение к данному интегралу ≈ -2.403, более точное, чем найденное выше при том же числе узлов $n=4$ и параметре $\alpha=0$.

К вычислению интегралов с бесконечной границей можно применять различные формулы численного интегрирования, пользуясь равенством, определяющим несобственный интеграл:

$$\int\limits_a^{\infty} f(x)dx := \lim_{b\to\infty}\int\limits_a^b f(x)dx\,.$$

Оно позволяет считать, что для достаточно больших значений b

$$\int\limits_a^{\infty} f(x)dx \approx \int\limits_a^b f(x)dx, \qquad (12.66)$$

и вычислять определенный интеграл $\int\limits_a^b f(x)dx$ с помощью известных квадратурных правил. Предполагая исходный интеграл абсолютно сходящимся, величину абсолютной погрешности, т.е. $\left|\int\limits_b^\infty f(x)dx\right|$, за счет увеличения b можно сделать сколь угодно малой. Реально это может быть достигнуто либо с помощью аналитического оценивания этой величины, либо таким устройством алгоритма вычисления данного несобственного интеграла на основе приближенного равенства (12.66), при котором граница b вычисляемого определенного интеграла постепенно перемещалась бы в процессе счета вправо по оси абсцисс до тех пор, пока величина интеграла не перестанет изменяться в требуемых десятичных знаках.

В случае несобственных интегралов типа В без ограничения общности можно считать, что подынтегральная функция имеет особенность на границе промежутка интегрирования, т.е. если точкой c, где $f(x)$ обращается в бесконечность, окажется внутренняя точка интервала (a, b), то данный интеграл можно представить символически как $\int\limits_a^c + \int\limits_c^b$. Также без потери общности, как мы уже знаем, достаточно рассматривать $\int\limits_{-1}^1 f(x)dx$. Но к таким интегралам, в которых подынтегральная функция имеет особыми точками -1 и (или) 1, можно применить квадратурную формулу Эрмита (Мелера) (12.61) в виде

$$\int\limits_{-1}^1 f(x)dx = \int\limits_{-1}^1 \frac{\sqrt{1-x^2}f(x)}{\sqrt{1-x^2}}dx \approx \frac{\pi}{n}\sum_{i=1}^n \sqrt{1-x_i^2}\, f(x_i) \qquad (12.67)$$

или более общую формулу (12.62), где параметры $\alpha > -1$, $\beta > -1$ желательно подобрать так, чтобы функция

$$\varphi(x) := (1-x)^{-\alpha}(1+x)^{-\beta} f(x)$$

была как можно более гладкой. Такой прием при вычислении несобственных интегралов называют ***мультипликативным выделением особенностей*** [19, 71, 78]. Существует несколько специальных квадратурных формул, позволяющих «загнать» в весовые функции различные типы особенностей: степенную, логарифмическую и др. [104].

Другой прием работы с интегралами типа В, предложенный Л.В.Канторовичем, называется ***аддитивным выделением особенностей***. Здесь, вообще говоря, нет необходимости заранее сводить интегрирование к стандартному промежутку. Суть приема состоит в том, что данный несобственный интеграл представляется в следующем виде:

$$\int\limits_a^b f(x)dx = \int\limits_a^b \varphi(x)dx + \int\limits_a^b [f(x)-\varphi(x)]dx, \qquad (12.68)$$

где функция $\varphi(x)$ подбирается так, чтобы она имела на $[a, b]$ такую же особенность, как и $f(x)$, и значение $\int\limits_a^b \varphi(x)dx$ находилось аналитическим путем, а функция $f(x)-\varphi(x)$ должна быть гладкой, т.е. интеграл $\int\limits_a^b [f(x)-\varphi(x)]dx$ без проблем должен вычисляться обычными квадратурами.

Для функций определенного класса со степенными особенностями разработана технология подбора таких функций $\varphi(x)$, которые принимают на себя особенность не только функции $f(x)$, но и некоторого числа ее производных, что позволяет ко второму интегралу в (12.68) применить квадратурные формулы заданного порядка [19, 61, 71]. Покажем, как это можно сделать.

Пусть $c \in [a, b]$ — особая точка функции $f(x)$ такой, что

$$f(x) = \frac{g(x)}{(x-c)^\gamma},$$

где $\gamma \in (0, 1)$, а $g(x) \in C^{k+1}[a, b]$, и пусть для вычисления второго интеграла в (12.68) предполагается использовать квадратурную формулу, требующую от функции $f(x)-\varphi(x)$ непрерывность k-ой производной. Введем в рассмотрение функцию

$$g_{k+1}(x) := g(c) + g'(c)(x-c) + \ldots + \frac{g^{(k+1)}(c)}{(k+1)!}(x-c)^{k+1}.$$

Так как эта функция есть отрезок разложения Тейлора функции $g(x)$ в окрестности особой точки $x = c$, то, полагая $\varphi(x) := \dfrac{g_{k+1}(x)}{(x-c)^\gamma}$, будем иметь равенство

$$f(x) - \varphi(x) = \frac{g(x) - g_{k+1}(x)}{(x-c)^\gamma} = o\left[(x-c)^k\right],$$

означающее непрерывную дифференцируемость функции $f(x) - \varphi(x)$ не менее k раз.

Имеются обобщения рассмотренного преобразования подынтегральной функции, например, учитывающие особенности логарифмического типа [19].

Пример 12.3. Дан интеграл $I := \int_0^1 \frac{e^x}{\sqrt{x}}\,dx$. Его подынтегральная функция $f(x) := \dfrac{e^x}{\sqrt{x}}$ имеет степенную особенность $\left(\gamma = \dfrac{1}{2}\right)$ в точке $x = 0$ и удовлетворяет требуемым выше условиям. Используя известное разложение

$$e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \ldots,$$

положим $g_3(x) := 1 + x + \dfrac{x^2}{2} + \dfrac{x^3}{6}$. Тогда

$$\varphi(x) := \frac{g_3(x)}{x^{\frac{1}{2}}} = x^{-\frac{1}{2}} + x^{\frac{1}{2}} + \frac{1}{2}x^{\frac{3}{2}} + \frac{1}{6}x^{\frac{5}{2}},$$

а

$$f(x) - \varphi(x) = \frac{6e^x - \left(6 + 6x + 3x^2 + x^3\right)}{6\sqrt{x}} \left(= \frac{1}{4!}x^{\frac{7}{2}} + \frac{1}{5!}x^{\frac{9}{2}} + \ldots\right). \qquad (12.69)$$

Согласно (12.68), непосредственно находим значение

$$I_1 := \int_0^1 \varphi(x)\,dx = \left[2\sqrt{x} + \frac{2}{3}x\sqrt{x} + \frac{1}{5}x^2\sqrt{x} + \frac{1}{21}x^3\sqrt{x}\right]_0^1 = 2\frac{32}{35}\ (\approx 2.914286),$$

а к интегралу $I_2 := \int_0^1 [f(x) - \varphi(x)]\,dx$ с полным основанием можно приме-

нить, например, эффективный алгоритм прямоугольников-трапеций (см. § 12.5), поскольку функция $f(x) - \varphi(x)$ имеет, по меньшей мере, непрерывную вторую производную. При этом из последнего представления этой функции в (12.69) легко видеть, что $f(0) - \varphi(0) = 0$. Подсчет значения $\int_0^1 [f(x) - \varphi(x)]\,dx$ указанным алгоритмом с заданной точностью $\varepsilon = 10^{-6}$ дает $I_2 = 0.011018$ (за 65 вычислений подынтегральной функции), что приводит к значению $I = I_1 + I_2 = 2.925304$.

Часто требуемый эффект аддитивного выделения особенности достигается интегрированием «по частям».

Так, в данном примере такой прием оказывается более простым:

$$\int_0^1 \frac{e^x}{\sqrt{x}}\,dx = 2\int_0^1 e^x\,d\sqrt{x} = 2\sqrt{x}e^x\Big|_0^1 - 2\int_0^1 \sqrt{x}e^x\,dx = 2e - 2I_3,$$

где $I_3 := \int_0^1 \sqrt{x}e^x\,dx$ — определенный интеграл с «хорошей» подынтегральной функцией. Применение того же алгоритма численного интегрирования с той же заданной точностью приводит к значению $I_3 = 1.255630$ (за 2049 вычислений функции). В итоге получаем то же приближенное значение $I = 2e - 2I_3 = 2.925304$.

Мультипликативным выделением особенности в этом примере можно воспользоваться напрямую, если применить специальную квадратурную формулу вида

$$\int_0^1 x^\alpha g(x)\,dx \approx \sum_{i=1}^n A_i g(x_i) \qquad (12.70)$$

при $\alpha = -0.5$ [104]:

$$\int_0^1 \frac{e^x}{\sqrt{x}}\,dx \approx \sum_{i=1}^n A_i e^{x_i}.$$

Например, при $n = 4$ с помощью таблицы узлов и весов формулы (12.70) из [104] находим

$$I \approx 0.725368 \cdot e^{0.033648} + 0.627413 \cdot e^{0.276184} + 0.444762 \cdot e^{0.634677} + 0.202457 \cdot e^{0.922157} \approx 2.925302,$$

а при $n = 6$ имеем

$$I \approx 0.498294 \cdot e^{0.015683} + 0.466985 \cdot e^{0.135300} + 0.406335 \cdot e^{0.344942} + 0.320157 \cdot e^{0.592750} + 0.213879 \cdot e^{0.817428} + 0.0943507 \cdot e^{0.963461} \approx 2.925304.$$

Налицо эффективность использования специальной квадратуры.

Применяются и иные приемы вычисления несобственных интегралов [13, 19, 78, 104, 149 и др.]. Заметим, что иногда достаточно сделать удачную замену переменной, чтобы преобразовать несобственный интеграл к более подходящему для вычисления виду.

УПРАЖНЕНИЯ

12.1. Выведите формулы остаточных членов для квадратурных формул левых и правых прямоугольников.

12.2. Сколько требуется знать значений подынтегральной функции для подсчета интеграла $\int_1^2 \frac{\ln x}{x} dx$ по формуле трапеций с точностью $\varepsilon = 0.01$?

12.3. А) Докажите, что если стоящая под знаком интеграла $I := \int_a^b f(x)dx$ функция $f(x)$ имеет знакопостоянную на (a,b) вторую производную, то справедлива оценка

$$\left| I - \frac{I^{\Pi} + I^{T}}{2} \right| \le \frac{1}{2} \left| I^{\Pi} - I^{T} \right|,$$

где I^{Π} и I^{T} — приближения к I, получаемые по формулам прямоугольников (12.8) и трапеций (12.28).

Каков содержательный смысл доказанного неравенства?

Б) Можно ли утверждать, что значение интеграла $\int_0^{\pi/3} \ln\cos x\, dx$ заключено между его приближенными значениями I^{Π} и I^{T}?

12.4. Найдите значения коэффициентов квадратурной формулы

$$\int_{-h}^{h} f(x)dx \approx Af(-h) + Bf(0) + Cf(h),$$

считая ее точной для многочленов второй степени; убедитесь, что она совпадает с формулой Симпсона.

12.5. Определите, с какой точностью можно вычислить $\int_0^1 \sin(e^x)dx$, привлекая девять значений подынтегральной функции:

а) по формуле прямоугольников;
б) по формуле трапеций;
в) по формуле Симпсона.

12.6. Покажите, что содержащаяся в выражении (12.8) при $n = 1$ простейшая формула средней точки является частным случаем квадратурной формулы Гаусса.

12.7. Непосредственным рассмотрением систем (12.47) и (12.49) убедитесь в совпадении формул Чебышева и Гаусса при $n = 2$ и найдите точные значения узлов t_1, t_2.

12.8. Рассмотрите разные частные случаи формулы Гаусса–Кристоффеля при $n = 1$ (квадратуры с одним узлом).

12.9. Подсчитайте коэффициенты квадратурной формулы Эрмита (12.59) для случая двух узлов непосредственно по формуле (12.60) (иначе, обоснуйте квадратурную формулу (12.61) при $n = 2$).

12.10. Постройте простейшую квадратурную формулу неинтерполяционного типа, заменяя подынтегральную функцию линейной функцией наилучшего среднеквадратического приближения, определяемой по трем равноотстоящим узлам (см. функцию $\varphi_i(x)$ при построении линейного фильтра в § 11.1). На основе полученной простейшей формулы запишите составную формулу такого типа.

ГЛАВА 13 || АППРОКСИМАЦИЯ ПРОИЗВОДНЫХ

Посредством дифференцирования интерполяционных формул выводятся формулы для приближенного вычисления производных. Остаточные члены полученных приближенных формул находятся с помощью формулы Тейлора и дифференцированием соответствующих остаточных членов интерполяционных формул. Особая роль придается простейшим аппроксимациям первого и второго порядков точности первой и второй производных в узлах сетки, что существенно используется в последующих главах. Обращается внимание на существование таких шагов аппроксимации производных по формулам различных порядков точности, при которых ограниченная точность вычисленных значений функции наименьшим образом влияет на точность результата.

13.1. ВЫВОД ФОРМУЛ ЧИСЛЕННОГО ДИФФЕРЕНЦИРОВАНИЯ

Численное дифференцирование, т.е. нахождение значений производных заданной функции $y = f(x)$ в заданных точках x, в отличие от рассмотренного в предыдущей главе численного интегрирования, можно считать не столь актуальной задачей в связи с отсутствием принципиальных трудностей с аналитическим нахождением производных. Однако имеется ряд моментов, не позволяющих обходить эту задачу стороной. Это и типичное для прикладных задач незнание аналитического вида $f(x)$, и возможное сильное усложнение функции при ее аналитическом дифференцировании (что затрудняет нахождение ее значений с высокой точностью), и желательность получения значений производных с помощью однотипных вычислительных процессов без привлечения аналитических выкладок. Главным же для дальнейшего является потребность в простых формулах, с помощью которых производные в заданных точках можно аппроксимировать несколькими значениями функции (быть может неизвестной) в этих и близких к ним точках.

Источником формул численного дифференцирования, как и большинства квадратурных формул, является полиномиальная интерполяция.

Зная в точках $x_i = x_0 + ih$ $(i = 0, 1, \ldots, n)$ при некотором $h > 0$ значения $y_i = f(x_i)$ данной функции $y = f(x)$, можно найти конечные разности $\Delta^k y_i$ и записать для нее, например, первый интерполяционный многочлен Ньютона $P_n(x)$ (см. (8.25)).

Дифференцируя приближенное равенство $f(x) \approx P_n(x)$, будем строить формулы приближенного дифференцирования разной точности в зависимости от степени n используемого интерполяционного многочлена. Через вспомогательную переменную $q := \dfrac{x - x_0}{h}$ приближенное представление функции $f(x)$ по первой формуле Ньютона выглядит наиболее просто:

$$f(x) \approx P_n(x_0 + qh) = y_0 + q\Delta y_0 + \frac{q(q-1)}{2!}\Delta^2 y_0 + \\ + \frac{q(q-1)(q-2)}{3!}\Delta^3 y_0 + \ldots + \frac{q(q-1)\ldots(q-n+1)}{n!}\Delta^n y_0 \quad (13.1)$$

(см. (8.26)). Отсюда получаем ***конечноразностную формулу численного дифференцирования***

$$f'(x) \approx q'_x [P_n(x_0 + qh)]'_q = \frac{1}{h}\left(y_0 + q\Delta y_0 + \frac{q^2 - q}{2}\Delta^2 y_0 + \right. \\ \left. + \frac{q^3 - 3q^2 + 2q}{6}\Delta^3 y_0 + \frac{q^4 - 6q^3 + 11q^2 - 6q}{24}\Delta^4 y_0 + \ldots \right)'_q ,$$

т.е.

$$f'(x) \approx \frac{1}{h}\left(\Delta y_0 + \frac{2q-1}{2}\Delta^2 y_0 + \frac{3q^2 - 6q + 2}{6}\Delta^3 y_0 + \right. \\ \left. + \frac{4q^3 - 18q^2 + 22q - 6}{24}\Delta^4 y_0 + \ldots \right). \quad (13.2)$$

При использовании последнего равенства для приближенного вычисления производной функции $f(x)$ в заданной точке $\tilde{x}$ из некоторой окрестности точки x_0 следует найти соответствующее значение $\tilde{q}$ переменной $q = \dfrac{x - x_0}{h}$ и подставить его в формулу (13.2). Максимальный порядок конечных разностей в этой формуле при желании получить производную с наибольшей точностью определяется в конкретной ситуации в соответствии с приведенными в §§ 8.4, 8.6 соображениями о выборе подходящей степени интерполяционного многочлена.

Аналогично можно вывести ряд других формул численного дифференцирования на основе различных интерполяционных формул (см. гл.8), более эффективно «работающих» вблизи других узловых точек и в общем случае не обязательно равноот-

стоящих. Предоставим читателю сделать это самостоятельно [19 и др.] и вернемся к приближенной формуле (13.2). Рассмотрим несколько ее частных случаев, фиксируя степень n лежащего в ее основе интерполяционного многочлена (13.1), равной 1, 2, 3. Этим значениям n отвечают соответственно одно, два, три первых слагаемых в формуле (13.2). Таким образом, имеем:

на основе линейной интерполяции

$$f'(x) \approx \frac{\Delta y_0}{h} \quad \text{для} \quad x \in (x_0 - \delta,\, x_1 + \delta), \tag{13.3}$$

на основе квадратичной интерполяции

$$f'(x) \approx \frac{1}{h}\left(\Delta y_0 + \frac{2q-1}{2}\Delta^2 y_0\right) \quad \text{для} \quad x \in (x_0 - \delta,\, x_2 + \delta), \tag{13.4}$$

на основе кубической интерполяции

$$f'(x) \approx \frac{1}{h}\left(\Delta y_0 + \frac{2q-1}{2}\Delta^2 y_0 + \frac{3q^2 - 6q + 2}{6}\Delta^3 y_0\right) \quad \text{для} \quad x \in (x_0 - \delta,\, x_3 + \delta), \tag{13.5}$$

и т.д. (при некоторых $\delta > 0$, определяющих промежуток экстраполяции приемлемого качества соответствующей интерполяционной формулой).

Для дальнейшего особый интерес представляют частные случаи формул (13.3)–(13.5), связывающие приближенное значение производной функции $f(x)$ в узлах $x_0, x_1, \ldots$ с узловыми значениями самой функции. Учитывая, что точкам x_0, x_1, x_2, x_3 соответствуют значения $q = 0, 1, 2, 3$, и раскрывая конечные разности через значения y_i $(i = 0, 1, 2, 3)$, имеем:

при $n = 1$ из (13.3)

$$f'(x_0) \approx y_0' := \frac{y_1 - y_0}{h}, \tag{13.6}$$

$$f'(x_1) \approx y_1' := \frac{y_1 - y_0}{h}; \tag{13.7}$$

при $n = 2$ из (13.4)

$$f'(x_0) \approx y_0' := \frac{1}{2h}(-3y_0 + 4y_1 - y_2), \tag{13.8}$$

$$f'(x_1) \approx y_1' := \frac{1}{2h}(y_2 - y_0), \tag{13.9}$$

$$f'(x_2) \approx y_2' := \frac{1}{2h}(y_0 - 4y_1 + 3y_2); \tag{13.10}$$

при $n = 3$ из (13.5)

$$f'(x_0) \approx y_0' := \frac{1}{6h}(-11y_0 + 18y_1 - 9y_2 + 2y_3),$$

$$f'(x_1) \approx y_1' := \frac{1}{6h}(-2y_0 - 3y_1 + 6y_2 - y_3),$$

$$f'(x_2) \approx y_2' := \frac{1}{6h}(y_0 - 6y_1 + 3y_2 + 2y_3),$$

$$f'(x_3) \approx y_3' := \frac{1}{6h}(-2y_0 + 9y_1 - 18y_2 + 11y_3).$$

В случае необходимости этот ряд формул можно продолжить с помощью общей формулы (13.2) или посмотреть в других источниках [19 и др.].

Повторное дифференцирование приближенного равенства (13.1), т.е. взятие производной по x от правой части формулы (13.2) с учетом $\frac{dq}{dx} = \frac{1}{h}$, приводит к конечноразностной формуле вычисления второй производной

$$f''(x) \approx \frac{1}{h^2}\left[\Delta^2 y_0 + (q-1)\Delta^3 y_0 + \frac{1}{12}\left(6q^2 - 18q + 11\right)\Delta^4 y_0 + \ldots\right], \tag{13.11}$$

из которой таким же образом следует приближенная формула для третьей производной

$$f'''(x) \approx \frac{1}{h^3}\left[\Delta^3 y_0 + \left(q - \frac{3}{2}\right)\Delta^4 y_0 + \ldots\right],$$

и т.д.

Наиболее важной в приложениях является простейшая аппроксимация второй производной с помощью постоянной $\frac{\Delta^2 y_0}{h^2}$ на промежутке $(x_0 - \delta,\, x_2 + \delta)$, получающаяся из (13.11) фиксированием только одного слагаемого (случай $n = 2$). В частности, в точке x_1 имеем приближенное равенство

$$f''(x_1) \approx y_1'' := \frac{y_0 - 2y_1 + y_2}{h^2}, \tag{13.12}$$

которое вместе с формулами (13.6)–(13.10) широко используется при построении конечноразностных методов решения краевых задач для обыкновенных дифференциальных уравнений второго порядка (см. гл. 17) и для уравнений в частных производных (гл. 19–21).

13.2. ОСТАТОЧНЫЕ ЧЛЕНЫ ПРОСТЕЙШИХ ФОРМУЛ ЧИСЛЕННОГО ДИФФЕРЕНЦИРОВАНИЯ

Чтобы получить представление о точности простейших аппроксимаций значений производных в узловых точках, определяемых формулами (13.6)–(13.10), (13.12), будем предполагать, что данная функция $f(x)$ обладает достаточной для выведения остаточных членов гладкостью. Кроме того, проведем в указанных формулах смещение индексов, т.е. будем считать, что исходная информация о функции соответствует изображенной на рис. 13.1 , и речь идет об аппроксимации производных в i-м узле x_i и/или в отстоящих от него на расстоянии h узлах x_{i-1} и x_{i+1}.

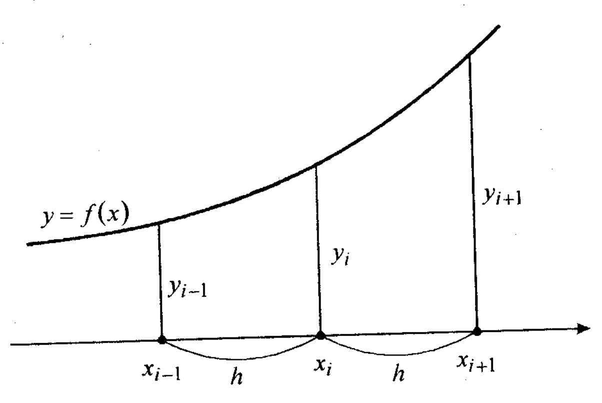


Рис. 13.1. *К простейшей аппроксимации производных*

Знание структуры приближенных выражений для производных, полученных из интерполяционных соображений, позволяет без особого труда (по крайней мере, для симметричных аппроксимаций) вывести формулы их остаточных членов, манипулируя разложениями $f(x)$ по формуле Тейлора подходящих порядков. Покажем это.

Простейшая несимметричная аппроксимация $f'(x_i)$ (формулы первого порядка точности). Запишем представление функции $f(x)$ по формуле Тейлора в окрестности точки x_i:

$$f(x) = f(x_i) + f'(x_i)(x - x_i) + \frac{f''(\xi)}{2}(x - x_i)^2.$$

Выразив отсюда $f'(x_i)$, имеем

$$f'(x_i) = \frac{f(x) - f(x_i)}{x - x_i} - \frac{f''(\xi)}{2}(x - x_i). \qquad (13.13)$$

Первый член правой части этого равенства — разностное отношение, аппроксимирующее производную вблизи x_i, а второй — остаточный член, характеризующий точность такой аппроксимации. При фиксировании в (13.13) $x = x_{i-1}$ одновременно зафиксируется и неизвестная точка $\xi = \xi_{i-1} \in (x_{i-1}, x_i)$; таким образом, приходим к *формуле левой аппроксимации* $f'(x_i)$ *с остаточным членом*:

$$f'(x_i) = \frac{y_i - y_{i-1}}{h} + \frac{f''(\xi_{i-1})}{2}h. \qquad (13.14)$$

Аналогично при $x = x_{i+1}$ из (13.13) получаем *формулу правой аппроксимации* $f'(x_i)$ *с остаточным членом*:

$$f'(x_i) = \frac{y_{i+1} - y_i}{h} - \frac{f''(\xi_{i+1})}{2}h. \qquad (13.15)$$

В приближенных равенствах

$$f'(x_i) \approx \frac{y_{i+1} - y_i}{h} \qquad (13.16)$$

при $i = 0$ и

$$f'(x_i) \approx \frac{y_i - y_{i-1}}{h} \qquad (13.17)$$

при $i = 1$ узнаём выведенные ранее формулы (13.6), (13.7), а остаточные члены в (13.14), (13.15) указывают на то, что, пользуясь аппроксимациями (13.16), (13.17), мы совершаем ошибку $O(h)$, т.е. эти формулы имеют первый порядок точности. Определенную информацию об ошибках левой и правой аппроксимаций первого порядка дает знание знаков остаточных членов.

Простейшая симметричная аппроксимация $f'(x_i)$ (формула второго порядка точности). Из разложения

$$f(x) = f(x_i) + f'(x_i)(x - x_i) + \frac{1}{2}f''(x_i)(x - x_i)^2 + \frac{1}{6}f'''(\xi)(x - x_i)^3$$

при $x = x_{i+1}$ и $x = x_{i-1}$ имеем соответственно

$$f(x_{i+1}) = f(x_i) + f'(x_i)h + \frac{1}{2}f''(x_i)h^2 + \frac{1}{6}f'''(\xi_{i+1})h^3$$

и

$$f(x_{i-1}) = f(x_i) - f'(x_i)h + \frac{1}{2}f''(x_i)h^2 - \frac{1}{6}f'''(\xi_{i-1})h^3.$$

Выполнив почленное вычитание двух последних равенств, получаем

$$y_{i+1} - y_{i-1} = 2hf'(x_i) + \frac{h^3}{6}[f'''(\xi_{i+1}) + f'''(\xi_{i-1})],$$

откуда с помощью теоремы о среднем, примененной к сумме третьих производных в квадратных скобках, приходим к *формуле симметричной аппроксимации* $f'(x_i)$ *с остаточным членом*:

$$f'(x_i) = \frac{y_{i+1} - y_{i-1}}{2h} - \frac{h^2}{6} f'''(\xi_i), \tag{13.18}$$

где ξ_i — некоторая точка интервала (x_{i-1}, x_{i+1}).

«Основная» часть формулы (13.18) —

$$f'(x_i) \approx \frac{y_{i+1} - y_{i-1}}{2h} \tag{13.19}$$

— при $i = 1$ совпадает с (13.9), а вид ее остаточного члена $-\frac{h^2}{6} f'''(\xi_i)$ означает, что аппроксимация (13.19) имеет второй порядок точности относительно шага h.

Простейшие аппроксимации второй производной. Из представления

$$f(x) = f(x_i) + f'(x_i)(x - x_i) + \frac{f''(x_i)}{2}(x - x_i)^2 + \\ + \frac{f'''(x_i)}{6}(x - x_i)^3 + \frac{f^{IV}(\xi)}{24}(x - x_i)^4$$

имеем

$$f(x_{i+1}) = f(x_i) + hf'(x_i) + \frac{h^2 f''(x_i)}{2} + \frac{h^3 f'''(x_i)}{6} + \frac{h^4 f^{IV}(\xi_{i+1})}{24},$$

$$f(x_{i-1}) = f(x_i) - hf'(x_i) + \frac{h^2 f''(x_i)}{2} - \frac{h^3 f'''(x_i)}{6} + \frac{h^4 f^{IV}(\xi_{i-1})}{24},$$

откуда почленным сложением получаем

$$y_{i+1} + y_{i-1} = 2y_i + h^2 f''(x_i) + \frac{h^4}{24}\left[f^{IV}(\xi_{i+1}) + f^{IV}(\xi_{i-1})\right].$$

Выражая из последнего равенства $f''(x_i)$, приходим к *формуле симметричной аппроксимации* $f''(x_i)$ *с остаточным членом*:

$$f''(x_i) = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} - \frac{h^2}{12} f^{IV}(\xi_i). \tag{13.20}$$

Остаточный член этой формулы $-\frac{h^2}{12} f^{IV}(\xi_i)$ с некоторым $\xi_i \in (x_{i-1}, x_{i+1})$ характеризует приближенное равенство

$$f''(x_i) \approx \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} \tag{13.21}$$

как аппроксимацию второй производной в точке x_i в т о р о г о порядка точности, т.е. с погрешностью $O(h^2)$.

То же отношение $\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2}$, используемое в качестве несимметричной аппроксимации второй производной функции $f(x)$, т.е. для вычисления приближенных значений $f''(x_{i-1})$ и $f''(x_{i+1})$, дает лишь первый порядок точности. Действительно, по формуле Тейлора для второй производной имеем

$$f''(x) = f''(x_i) + f'''(\xi)(x - x_i).$$

Подставляя сюда вместо $f''(x_i)$ правую часть равенства (13.20), получаем

$$f''(x) = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} + f'''(\xi)(x - x_i) - \frac{h^2}{12} f^{IV}(\xi_i).$$

Из этого равенства при $x = x_{i+1}$ и $x = x_{i-1}$ следуют частные *формулы несимметричной аппроксимации второй производной с остаточными членами*:

$$f''(x_{i-1}) = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} - hf'''(\xi_{i-1}) - \frac{h^2}{12} f^{IV}(\xi_i)$$

и

$$f''(x_{i+1}) = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} + hf'''(\xi_{i+1}) - \frac{h^2}{12} f^{IV}(\xi_i).$$

С помощью формулы Тейлора можно вывести остаточные члены и других простейших аппроксимаций производных.

Более общий подход для получения выражений остаточных членов интерполяционных формул численного дифференцирования состоит в дифференцировании остаточного члена интерполяционной формулы Лагранжа [19, 61]. Согласно (8.13),

$$R_n(x) := f(x) - L_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\Pi_{n+1}(x), \tag{13.22}$$

где $L_n(x)$ — интерполяционный многочлен Лагранжа, построенный для $n+1$ раз дифференцируемой функции $f(x)$ по $n+1$ узлу $x_0, x_1, \ldots, x_n$ (неважно, в какой форме), ξ — некоторая точка из интервала (x_0, x_n), а $\Pi_{n+1}(x) := (x-x_0)(x-x_1)\ldots(x-x_n)$. Из (13.22) следует

$$R_n'(x) := f'(x) - L_n'(x) =$$

$$= \frac{1}{(n+1)!}\left\{\frac{d}{dx}\left[f^{(n+1)}(\xi)\right]\Pi_{n+1}(x) + f^{(n+1)}(\xi)\frac{d}{dx}[\Pi_{n+1}(x)]\right\}. \quad (13.23)$$

Если величина $\frac{d}{dx}\left[f^{(n+1)}(\xi)\right]$ ограничена, то при подстановке в последнее выражение узловых значений $x = x_i \quad (i = 0, 1, \ldots, n)$ за счет $\Pi_{n+1}(x_i) = 0$ получим простую формулу остаточного члена аппроксимаций $f'(x_i) \approx L'(x_i)$ первой производной в узлах интерполяции:

$$f'(x_i) - L'(x_i) = \frac{f^{(n+1)}(\xi_i)}{(n+1)!}\Pi_{n+1}'(x_i), \quad \xi_i \in (x_0, x_n). \quad (13.24)$$

В случае равноотстоящих узлов $x_i = x_0 + ih$, который здесь, в основном, и рассматривается,

$$\Pi_{n+1}'(x_i) = (-1)^{n-i}\, i!\,(n-i)!\,h^n,$$

вследствие чего равенство (13.24) трансформируется в формулу

$$f'(x_i) - y_i' = (-1)^{n-i}\frac{i!(n-i)!}{(n+1)!}h^n f^{(n+1)}(\xi_i). \quad (13.25)$$

Из нее отчетливо видно, что при аппроксимации первой производной в точках x_i значениями $y_i' := L_n'(x_i)$, получаемыми дифференцированием интерполяционного многочлена n-й степени, остаточный член имеет n-й порядок относительно шага аппроксимации h.

Воспользуемся формулой (13.25), чтобы выписать остаточные члены несимметричных аппроксимаций первой производной (13.8) и (13.10). При $n = 2$, $i = 0$ и $i = 2$ из (13.25) следует

$$f'(x_0) - y_0' = \frac{h^2}{3}f'''(\xi_0)$$

и

$$f'(x_2) - y_2' = \frac{h^2}{3}f'''(\xi_2).$$

Ставя эти формулы в один ряд с формулой (13.18), имеем:

$$f'(x_{i-1}) = \frac{-3y_{i-1} + 4y_i - y_{i+1}}{2h} + \frac{h^2}{3}f'''(\xi_{i-1}), \quad (13.26)$$

$$f'(x_{i+1}) = \frac{y_{i-1} - 4y_i + 3y_{i+1}}{2h} + \frac{h^2}{3}f'''(\xi_{i+1}). \quad (13.27)$$

Заметим, что *формулы несимметричной аппроксимации* $f'(x_{i-1})$ (13.26) *и* $f'(x_{i+1})$ (13.27) *второго порядка точности* имеют в остаточном члене вдвое больший коэффициент, чем формула симметричной аппроксимации (13.18).

Для оценивания погрешности численного дифференцирования при значениях аргумента, не совпадающих с узловыми, и для получения остаточных членов приближенных формул $f^{(k)}(x) \approx L_n^{(k)}(x)$ при $k > 1$ формула (13.23) малопригодна. В книге [19] на основе интерполяционной формулы Ньютона для неравных промежутков (8.43) и связей между разделенными разностями и производными выводятся следующие формулы остаточных членов:

$$f'(x) - L_n'(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\Pi_{n+1}'(x) + \frac{f^{(n+2)}(\xi_1)}{(n+2)!}\Pi_{n+1}(x), \quad (13.28)$$

$$f''(x) - L_n''(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\Pi_{n+1}''(x) +$$

$$+ 2\frac{f^{(n+2)}(\xi_1)}{(n+2)!}\Pi_{n+1}'(x) + 2\frac{f^{(n+3)}(\xi_2)}{(n+3)!}\Pi_{n+1}(x), \quad (13.29)$$

и вообще, для $k \in \{0, 1, \ldots, n\}$

$$f^{(k)}(x) - L_n^{(k)}(x) =$$

$$= \sum_{j=0}^{k}\frac{k!}{(k-j)!(n+j+1)!}f^{(n+j+1)}(\xi_j)\Pi_{n+1}^{(k-j)}(x). \quad (13.30)$$

Имеются и другие, отличные от (13.28)–(13.30) представления остаточных членов и оценок погрешностей интерполяционных формул численного дифференцирования [44, 101 и др.].

Так как остаточные члены формул численного дифференцирования выражаются через производные более высоких порядков, чем искомые производные, то «в чистом виде» подобные выражения остаточных членов малопригодны для практического оценивания погрешностей. Однако не стоит считать их совсем

бесполезными. Во-первых, они важны для качественного сравнения (например, по порядку) различных аппроксимаций производных. Во-вторых, их можно применить для оценивания точности результатов численного дифференцирования, пользуясь простейшими связями между производными и конечными (или разделенными) разностями, например, составляя таблицы конечных разностей до требуемого в оценке порядка и принимая за $\max\left|f^{(n+1)}(x)\right|$ величину $\max\left\{\left|\Delta^{n+1}y_i\right|\right\}/h^{n+1}$ в соответствующей области таблицы. В третьих, выполнение условий, при которых справедливы те или иные формулы аппроксимации производных на равномерной сетке, позволяет при вычислении значений производных применять принцип Рунге двойного счета с разными шагами подобно тому, как это делалось в § 12.5 при вычислении интегралов (более подробно об этом см. в [78, 101]).

Пример 13.1. Бесконечно гладкая функция $y = f(x)$ задана таблицей своих значений и значений конечных разностей (табл.13.1).

Таблица 13.1

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$	$\Delta^4 y$	$\Delta^5 y$
1.0	0.367879					
		−0.035008				
1.1	0.332871		0.003331			
		−0.031677		−0.000316		
1.2	0.301194		0.003015		0.000028	
		−0.028662		−0.000288		0.000001
1.3	0.272532		0.002727		0.000029	
		−0.025935		−0.000259		−0.000004
1.4	0.246597		0.002468		0.000025	
		−0.023467		−0.000234		−0.000005
1.5	0.223130		0.002234		0.000020	
		−0.021233		−0.000214		0.000002
1.6	0.201897		0.002020		0.000022	
		−0.019213		−0.000192		−0.000003
1.7	0.182684		0.001828		0.000019	
		−0.017385		−0.000173		−0.000005
1.8	0.165299		0.001655		0.000014	
		−0.015730		−0.000159		
1.9	0.149569		0.001496			
		−0.014234				
2.0	0.135335					

Найдем приближения к значению $f'(1.5)$, пользуясь простейшими аппроксимациями по формулам первого и второго порядков точности. Имеем:

по формуле (13.16)

$$f'(1.5) \approx \frac{f(1.6) - f(1.5)}{0.1} \approx -0.21233;$$

по формуле (13.17)

$$f'(1.5) \approx \frac{f(1.5) - f(1.4)}{0.1} \approx -0.23467;$$

по формуле (13.19)

$$f'(1.5) \approx \frac{f(1.6) - f(1.4)}{0.2} \approx -0.22350; \qquad (13.31)$$

по формуле (13.26)

$$f'(1.5) \approx \frac{-3f(1.5) + 4f(1.6) - f(1.7)}{0.2} \approx -0.22243;$$

по формуле (13.27)

$$f'(1.5) \approx \frac{f(1.3) - 4f(1.4) + 3f(1.5)}{0.2} \approx -0.22233.$$

Если на основании данной таблицы конечных разностей посчитать, что

$$\max_{x\in[1,\,2]}\left|f''(x)\right| \approx \frac{\max\left|\Delta^2 y\right|}{h^2} \approx 0.3 \quad \text{и} \quad \max_{x\in[1,\,2]}\left|f'''(x)\right| \approx \frac{\max\left|\Delta^3 y\right|}{h^3} \approx 0.3,$$

то в соответствии с выведенными выше формулами остаточных членов находим следующие оценки абсолютных погрешностей для полученных приближенных значений $f'(1.5)$:

$\approx \frac{0.3}{2} \cdot 0.1 = 0.015$ для первых двух,

$\approx \frac{0.3}{6} \cdot 0.01 = 0.0005$ для третьего,

$\approx \frac{0.3}{3} \cdot 0.01 = 0.001$ для двух последних.

Откроем теперь факт, что исходная таблица составлена для функции e^{-x} и, значит, $f'(x) = -f(x)$. Это позволяет сравнить полученные приближенные значения $f'(1.5)$ с истинным $f'(1.5) = -f(1.5)$. Разница между приближенными значениями $f'(1.5)$ (без учета знака) составляет соответственно

$$0.01080\,,\quad 0.01114\,,\quad 0.00037\,,\quad 0.00070\,,\quad 0.00080$$

— числа, достаточно хорошо вписывающиеся в подсчитанные оценки

абсолютных погрешностей*).

Попытаемся при вычислении $f'(1.5)$ применить принцип Рунге. Считая, что невозможно пополнить таблицу новыми значениями функции, остается заложить процесс увеличения шага, т.е. включить механизм прореживания таблицы.

Примем за основу формулу симметричной аппроксимации второго порядка точности, а за начальный шаг — шаг $h=0.1$ исходной таблицы. Обозначив приближенное значение $f'(1.5)$ через $y'(h)$, в соответствии с правилом Рунге уточняем его по формуле

$$y'(h)+\frac{y'(h)-y'(2h)}{3}.$$

Учитывая, что $y'(h)\approx-0.22350$ (см. (13.31)), а

$$y'(2h)=\frac{f(1.7)-f(1.3)}{0.4}\approx-0.22462,$$

находим поправку Ричардсона

$$\frac{y'(h)-y'(2h)}{3}\approx 0.000373,$$

прибавление которой к $y'(h)$ дает значение $f'(1.5)\approx-0.223127$, лишь в последнем знаке отличающееся от точного. Модуль этой поправки служит приближенной оценкой погрешности значения $y'(h)$, как видим, неплохо согласующейся по порядку с полученной выше оценкой (0.0005).

Для $f''(1.5)$ по симметричной формуле (13.21) с шагом $h=0.1$ имеем

$$y''(0.1)=\frac{f(1.4)-2f(1.5)+f(1.6)}{0.01}\approx 0.2234,$$

с шагом $2h=0.2$ —

$$y''(0.2)=\frac{f(1.3)-2f(1.5)+f(1.7)}{0.04}\approx 0.2239.$$

Поправка Ричардсона равна

$$\frac{0.2234-0.2239}{3}\approx-0.00017;$$

ее прибавление к значению $y''(0.1)$ дает уточненное значение

$$f''(1.5)\approx 0.2234-0.00017=0.22323.$$

В соответствии с выражением остаточного члена в (13.20), принимая

$$\max_{x\in[1,\,2]}\left|f^{IV}(x)\right|\approx\frac{\max\left|\Delta^4 y\right|}{h^4}\approx 0.3,$$

*) Указанные оценки могут быть немного точнее, если максимумы модулей конечных разностей при оценивании модулей производных, входящих в остаточные члены, учитывать только в используемой части таблицы.

находим оценку погрешности значения $y''(0.1)$:

$$\left|f''(1.5)-y''(0.1)\right|\le\frac{0.3}{12}\cdot 0.01=0.00025;$$

близок к ней и модуль поправки Ричардсона. Однако истинная ошибка $0.2234-0.22313=0.00027$, хотя и имеет тот же порядок, но превосходит ту и другую оценки. Это можно объяснить влиянием погрешностей округления исходных данных, более заметным при приближенном вычислении старших производных (см. следующий параграф). Нетрудно видеть, что по содержащимся в таблице данным найти какие-либо значения пятой и последующих производных в принципе невозможно.

Замечание 13.1. Обратим внимание на одинаковость структуры остаточных членов формул численного дифференцирования и интегрирования, опирающихся на полиномиальное интерполирование данной функции по системе равноотстоящих узлов с шагом h. Вхождение в выражения остаточных членов старших производных, порядок которых определенным образом согласован с показателем степени шага, наводит на мысль о возможности и здесь, т.е. при численном дифференцировании, пользоваться понятием алгебраического порядка точности. Структура же самих формул приближенного дифференцирования, имеющих вид

$$f^{(k)}(x)\approx\sum_{i=1}^{n}c_i f(x_i),\quad k\in\mathbf{N}, \tag{13.32}$$

позволяет строить конкретные формулы такого типа методом неопределенных коэффициентов, подбирая коэффициенты c_i в (13.32) (и, возможно, узлы x_i) так, чтобы формула (13.32) была точна для произвольных многочленов некоторой фиксированной степени (например, наиболее высокой при заданном h) [12].

Так, желая получить приближенную формулу

$$f''(x)\approx c_1 f(x_i-h)+c_2 f(x_i)+c_3 f(x_i+h),$$

будем подставлять в нее вместо $f(x)$ последовательно степенные функции $1, x, x^2$, вторые производные которых при любом x равны соответственно 0, 0, 2. Получаем систему

$$\begin{cases} c_1+c_2+c_3=0,\\ c_1(x_i-h)+c_2x_i+c_3(x_i+h)=0,\\ c_1(x_i-h)^2+c_2x_i^2+c_3(x_i+h)^2=2. \end{cases}$$

Считая, что в качестве x_i может быть взята любая точка, полагаем $x_i = 0$, и из упрощенной таким образом системы

$$\begin{cases} c_1 + c_2 + c_3 = 0, \\ (c_3 - c_1)h = 0, \\ (c_1 + c_3)h^2 = 2 \end{cases}$$

находим $c_1 = c_3 = \frac{1}{h^2}$, $c_2 = -\frac{2}{h^2}$. В результате приходим к формуле

$$f''(x) = \frac{f(x_{i-1}) - 2f(x_i) + f(x_{i+1})}{h^2},$$

при $x \in [x_{i-1}, x_{i+1}]$ точной для многочленов второй степени, а при $x = x_i$, как нам уже известно (см. (13.20)), она точна для многочленов третьей степени.

13.3. ОПТИМИЗАЦИЯ ШАГА ЧИСЛЕННОГО ДИФФЕРЕНЦИРОВАНИЯ ПРИ ОГРАНИЧЕННОЙ ТОЧНОСТИ ЗНАЧЕНИЙ ФУНКЦИИ

Откажемся от использовавшегося ранее обозначения $y_i = f(x_i)$ и будем полагать, что y_i — это приближенное значение функции $f(x_i)$ в точке x_i. Предположим, что уровень абсолютных погрешностей значений y_i в разных узлах x_i примерно одинаков и ограничен числом $\delta > 0$. Таким образом, имеем

$$f(x_i) \approx y_i\,(\pm\delta), \quad \text{т.е.} \quad f(x_i) \in (y_i - \delta,\ y_i + \delta).$$

Будем изучать влияние неточностей задания (вычисления) значений функции на результаты аппроксимации производных по этим значениям для различных простейших формул.

Рассмотрим простейшую формулу (13.15) аппроксимации первой производной в узле x_i правым разностным отношением:

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{h} - \frac{h}{2} f''(\xi_{i+1}).$$

Так как при вычитании приближенных чисел их абсолютные погрешности складываются, то замена неизвестного точно разностного отношения $\frac{f(x_{i+1}) - f(x_i)}{h}$ реально вычисляемым отношением $\frac{y_{i+1} - y_i}{h}$ порождает ошибку, оцениваемую по модулю величиной $\frac{2\delta}{h}$. А поскольку погрешность аппроксимации $f'(x)$ точным отношением $\frac{f(x_{i+1}) - f(x_i)}{h}$ в соответствии с выражением остаточного члена можно оценить величиной $\frac{M_2}{2}h$, где $M_2 := \max\limits_{x \in [x_i,\, x_{i+1}]} |f''(x)|$, приходим к неравенству

$$\left| f'(x_i) - \frac{y_{i+1} - y_i}{h} \right| \le \frac{2\delta}{h} + \frac{M_2}{2} h.$$

Итак, полная абсолютная погрешность приближенного равенства

$$f'(x_i) \approx \frac{y_{i+1} - y_i}{h} \qquad (13.33)$$

оценивается величиной

$$g(h) := \frac{2\delta}{h} + \frac{M_2}{2} h, \qquad (13.34)$$

образованной двумя слагаемыми: вычислительной погрешностью, порождаемой неточными значениями функции, и погрешностью аппроксимации, связанной с выбором формулы численного дифференцирования. При $h \to 0$ за счет первого слагаемого функция $g(h)$ бесконечно растет; при сравнительно больших h с ростом h также будет наблюдаться рост $g(h)$ (асимптотически линейный), благодаря второму слагаемому. Схематично поведение графика функции $g = g(h)$ отображено на рис. 13.2.

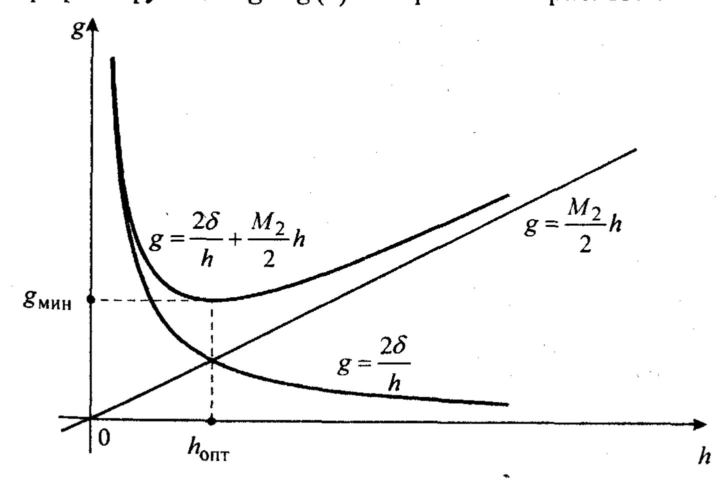


Рис. 13.2. Зависимость границы полной погрешности (13.34) приближенного равенства (13.33) от шага h

Не вызывает сомнений существование такого значения $h = h_{\text{опт}}$, при котором верхняя граница $g(h)$ полной погрешности приближенного равенства (13.33) минимальна. Это значение легко найти, приравняв нулю производную функции $g(h)$. Имеем:

$$-\frac{2\delta}{h^2} + \frac{M_2}{2} = 0,$$

откуда $h_{\text{опт}} = 2\sqrt{\dfrac{\delta}{M_2}}$ — ***оптимальный шаг***. Из этой формулы видно, что сделать аппроксимацию производной в точке x_i приближенным разностным отношением (13.33) сколь угодно качественной можно лишь в том случае, когда величину δ можно сделать сколь угодно малой (тогда можно рассчитывать на малость $h_{\text{опт}}$, что, в свою очередь, обеспечит малость $g_{\text{мин}} = g(h_{\text{опт}})$). Реально же при машинных вычислениях дело обстоит следующим образом [12].

В лучшем случае, значения функции по точности ограничены лишь разрядностью мантиссы ЭВМ. Тогда, грубо полагая $\delta = 2^{-m}$, $M_2 = 1$, находим, что $h_{\text{опт}} = 2 \cdot 2^{-\frac{m}{2}}$ и , следовательно.

$$g_{\text{мин}} = \frac{2 \cdot 2^{-m}}{2 \cdot 2^{-\frac{m}{2}}} + \frac{2 \cdot 2^{-\frac{m}{2}}}{2} = 2 \cdot 2^{-\frac{m}{2}}. \qquad (13.35)$$

Этот результат говорит о том, что *при простейших аппроксимациях производной с помощью правых разностных отношений (очевидно, для левых — то же самое) происходит потеря точности, составляющая половину значащих цифр числа.* Такой вывод не является для нас неожиданным (по крайней мере, качественно), поскольку потеря первых слева значащих цифр при вычитании близких значений функции при составлении таблиц конечных разностей с малым шагом уже обсуждалась в гл.8; здесь же эти разности еще делятся на малый шаг, что ведет к увеличению абсолютной погрешности приближенных значений производной.

Для формул численного дифференцирования более высоких порядков точности ситуация с потерей значащих разрядов в результатах их применения обстоит несколько лучше. Действительно, рассмотрим в тех же условиях, например, формулу симметричной аппроксимации второго порядка точности для $f'(x_i)$.

Полагая $M_3 := \max|f'''(x)|$, согласно (13.18), аналогично предыдущему имеем:

$$\left| f'(x_i) - \frac{y_{i+1} - y_{i-1}}{2h} \right| \le g(h) := \frac{\delta}{h} + \frac{M_3}{6} h^2,$$

откуда, решая уравнение

$$-\frac{\delta}{h^2} + \frac{M_3}{3} h = 0,$$

находим $h_{\text{опт}} = \sqrt[3]{\dfrac{3\delta}{M_3}}$. Теперь, если посчитать для простоты $M_3 = 3$, $\delta = 2^{-m}$, получим $h_{\text{опт}} = \delta^{\frac{1}{3}} = 2^{-\frac{m}{3}}$ и

$$g_{\text{мин}} = g(h_{\text{опт}}) = \frac{2^{-m}}{2^{-\frac{m}{3}}} + \frac{1}{2} \cdot 2^{-\frac{2}{3}m} = \frac{3}{2} \cdot 2^{-\frac{2}{3}m}.$$

Как видим, в этой ситуации можно рассчитывать на сохранение в приближенной производной примерно двух третей двоичных разрядов, имевшихся в исходных значениях функции.

Симметричная формула второго порядка точности для второй производной (13.20) в реальных условиях обеспечивает сохранение половины разрядов, как и рассмотренная выше формула первого порядка точности для первой производной. Действительно, с $\max|f^{IV}(x)| = M_4$ из (13.20) имеем оценку

$$\left| f''(x_i) - \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} \right| \le \frac{4\delta}{h^2} + \frac{M_4}{12} h^2 =: g(h),$$

наименьшее значение которой реализуется при $h_{\text{опт}} = 2 \cdot \sqrt[4]{\dfrac{3\delta}{M_4}}$.

При $M_4 = 3$, $\delta = 2^{-m}$ получаем $h_{\text{опт}} = 2 \cdot 2^{-\frac{m}{4}}$ и $g_{\text{опт}} = 2 \cdot 2^{-\frac{m}{2}}$ (сравните с формулой (13.35)).

Каждое последующее (повторное) дифференцирование влечет новые потери точности. Об этом можно судить, во-первых, по росту показателя степени h с ростом порядка производной в знаменателе формул численного дифференцирования, во-вторых, по пропаданию значащих цифр в конечных разностях увеличи-

вающихся порядков, лежащих в основе этих формул. Для получения приближенных значений удовлетворительной точности при вычислении производных достаточно высоких порядков может потребоваться огромный запас знаков в значениях функции, что трудно обеспечить на практике.

Замечание 13.2. [92] Посмотрим с той же точки зрения на формулы численного интегрирования. Пусть интеграл $I := \int_{-1}^{1} f(x)dx$ вычисляется с помощью какой-либо квадратурной формулы $I \approx \sum_i A_i y_i$, и пусть, как и выше, $f(x_i) \approx y_i\,(\pm\delta)$. Тогда, так как $\sum_i A_i = 2$ (в общем случае, при интегрировании на промежутке $[a, b]$ должно быть $\sum_i A_i = b - a$), то

$$\left|I - \sum_i A_i y_i\right| = \left|I - \sum_i A_i f(x_i) + \sum_i A_i f(x_i) - \sum_i A_i y_i\right| \le$$

$$\le \left|I - \sum_i A_i f(x_i)\right| + \sum_i A_i \mid f(x_i) - y_i \mid \le \left|I - \sum_i A_i f(x_i)\right| + 2\delta.$$

Последнее равенство характеризует *квадратурный процесс как численно устойчивый*: независимо от числа используемых для приближенного вычисления интеграла значений функции, содержащих ошибки в пределах интервала $(-\delta, \delta)$, связанная с этим ошибка результата оценивается величиной $O(\delta)$, т.е. того же порядка, а это позволяет заботиться лишь об обеспечении требований к малости $\varepsilon > 0$ погрешности применяемой квадратурной формулы (разумеется, при условии $\delta < \varepsilon$).

Невозможность сделать погрешность аппроксимации производных разностными отношениями сколь угодно малой даже при убывании погрешностей вычисления значений функции вынуждает считать задачу численного дифференцирования ***некорректной***. Здесь нарушается одно из условий корректности, а именно, нет непрерывной зависимости точности результата от точности входных данных в том смысле, что в процессе $h \to 0$ ошибка в значениях функции может убывать, а ошибка результата — бесконечно расти. Покажем, что такое возможно [78].

Пусть значение функции $f(x)$ в точке x представляется последовательностью приближений $y_k(x)$ с ошибками, уменьшающимися при $k \to \infty$ по закону $\frac{\sin(k^2 x)}{k}$, т.е.*)

$$f(x) = y_k(x) + \frac{\sin(k^2 x)}{k} \quad \forall k \in \mathbf{N}. \tag{13.36}$$

Тогда, если при использовании формулы правой аппроксимации производной

$$f'(x) \approx \frac{f(x+h) - f(x)}{h} \tag{13.37}$$

вместо $f(x)$ будут подставляться приближения $y_k(x)$, т.е. будет применяться формула

$$f'(x) \approx y'_k(x) := \frac{y_k(x+h) - y_k(x)}{h},$$

то порождаемая такой подменой ошибка на основе (13.36) при каждом $k \in N$ будет составлять величину

$$E := \frac{\sin\left(k^2(x+h)\right) - \sin\left(k^2 x\right)}{kh} =$$

$$= \frac{2\sin\left(\frac{k^2 h}{2}\right)\cos\left(k^2\left(x + \frac{h}{2}\right)\right)}{kh} = k \cdot \frac{\sin\frac{k^2 h}{2}}{\frac{k^2 h}{2}} \cdot \cos\left(k^2\left(x + \frac{h}{2}\right)\right).$$

Последнее представление показывает, что если $k \to \infty$ и $h \to 0$ так, что $k^2 h \to 0$, то величина ошибки E может неограниченно возрастать. Следовательно, $y'_k(x) \not\to f'(x)$.

Попытаемся убедиться в том, что задача численного дифференцирования на самом деле регуляризуема, и что, например, только что рассматривавшаяся формула (13.37) аппроксимации производной правым разностным отношением при определенных условиях на шаг h может считаться для нее регуляризирующим оператором в смысле определения 1.5 из § 1.7. С этой целью в терминах § 1.7 ставим задачу так: требуется найти приближенно

*) Обратим внимание, что здесь, в отличие от рассуждений, проводившихся в начале этого параграфа, мы следим за ошибками, а не за оценками их модулей.

значение $y = \dfrac{df(x)}{dx}$ в точке $x \in (a, b)$, считая, что вместо точных значений $f(x)$ при $x \in [a, b]$ известны их приближенные значения

$$\tilde{f}_\delta(x) := f(x) + \varphi(x)$$

такие, что

$$|\varphi(x)| \le \delta \quad \forall\, x \in [a, b]. \qquad (13.38)$$

В соответствии с приближенной формулой (13.37) положим

$$R(f, \alpha) := \frac{f(x+\alpha) - f(x)}{\alpha},$$

где α — положительный параметр, не выводящий $x+\alpha$ за пределы отрезка $[a, b]$. Тогда, согласно определению 1.5, α-регуляризованное решение поставленной задачи должно иметь вид

$$y_\alpha = R(\tilde{f}_\delta, \alpha(\delta)) = \frac{\tilde{f}_\delta(x+\alpha) - \tilde{f}(x)}{\alpha} =$$

$$= \frac{f(x+\alpha) - f(x)}{\alpha} + \frac{\varphi(x+\alpha) - \varphi(x)}{\alpha}.$$

Так как по определению производной

$$\frac{f(x+\alpha) - f(x)}{\alpha} \to y \quad \text{при} \quad \alpha \to 0,$$

а, в силу (13.38),

$$\left|\frac{\varphi(x+\alpha) - \varphi(x)}{\alpha}\right| \le \frac{2\delta}{\alpha},$$

то требуемая сходимость $y_\alpha \xrightarrow[\delta \to 0]{} y$ будет наблюдаться в том случае, если

$$\alpha \to 0 \quad \text{и} \quad \frac{2\delta}{\alpha} \to 0 \quad \text{при} \quad \delta \to 0. \qquad (13.39)$$

Чтобы удовлетворить (13.39), достаточно подобрать параметр $\alpha = \alpha(\delta)$ в виде $\alpha = \dfrac{\delta}{\eta(\delta)}$ так, чтобы величина $\eta(\delta)$ стремилась к нулю одновременно с δ и при этом было $\delta = o(\eta(\delta))$. В таком случае будем иметь

$$\alpha = \frac{\delta}{\eta(\delta)} \xrightarrow[\delta \to 0]{} 0 \quad \text{и} \quad \frac{2\delta}{\delta/\eta(\delta)} = 2\eta(\delta) \xrightarrow[\delta \to 0]{} 0,$$

т.е. выполнение условий (13.39).

Возьмем, например, $\eta(\delta) := \sqrt{\delta}$. Тогда $\alpha = \dfrac{\delta}{\sqrt{\delta}} = \sqrt{\delta}$, $\dfrac{2\delta}{\alpha} = \dfrac{2\delta}{\sqrt{\delta}} = 2\sqrt{\delta}$ и, следовательно,

$$\tilde{y}_{\sqrt{\delta}} = R(\tilde{f}_\delta, \sqrt{\delta}) = \frac{\tilde{f}_\delta(x+\sqrt{\delta}) - \tilde{f}_\delta(x)}{\sqrt{\delta}} \qquad (13.40)$$

— конкретное регуризованное решение поставленной задачи, дающее устойчивые приближения к производной $\dfrac{df(x)}{dx}$ при любом $\delta > 0$.

Продемонстрированный подход к проблеме численного дифференцирования при неточных данных и его результат (13.40) делают мотивированными рекомендации по выбору шага h аппроксимации первых производных по формулам первого порядка точности типа $h = O(\sqrt{\delta})$, где δ — уровень погрешностей вычисления значений функции (или $h = \sqrt{masheps}$, если есть основание считать $\delta \approx masheps$).

УПРАЖНЕНИЯ

13.1. Выведите частные формулы аппроксимации второй производной в равноотстоящих узлах, основываясь на кубической интерполяции (например, из общей конечноразностной формулы (13.11)).

13.2. Запишите симметричную формулу четвертого порядка точности для аппроксимации $f'(x_i)$ и выведите ее остаточный член, используя формулу Тейлора.

13.3. Убедитесь, что выведенные в § 13.2 с помощью формулы Тейлора выражения остаточных членов простейших аппроксимаций первой и второй производных являются частными случаями общих формул (13.28), (13.29).

13.4. Выведите общую формулу численного дифференцирования, привлекая интерполяционную формулу Ньютона для неравных промежутков (8.43).

13.5. Непосредственной подстановкой убедитесь, что приближенная формула

$$f''(0) \approx \frac{f(-h) - 2f(0) + f(h)}{h^2}$$

точна для произвольного многочлена второй степени.

13.6. Запишите формулу полной погрешности и найдите выражение оптимального шага таблицы значений функции $f(x)$ при аппроксимации первой производной по формуле несимметричной аппроксимации второго порядка точности, считая, что абсолютные погрешности значений функции в узлах не превосходят δ и что $\max|f'''(x)| \le M_3$.

13.7. Укажите значения оптимального шага и наивысшую гарантированную точность аппроксимации первой производной функции $y = \sin x$ посредством вычисляемых с предельной абсолютной погрешностью $\delta = 10^{-6}$ ее значений по различным простейшим формулам первого и второго порядков точности.

13.8. Найдите величины оптимального шага и оцените наилучшую точность, которая может быть гарантирована в условиях примера 13.1 (см. § 13.2) для различных рассмотренных там аппроксимаций первой и второй производных.

13.9. Постройте простейший регуляризатор для задачи численного нахождения $y = \dfrac{d^2 f(x)}{dx^2}$ в условиях неточного вычисления значений $f(x)$ (с известным уровнем δ абсолютных погрешностей).

ГЛАВА 14 МЕТОДЫ ЭЙЛЕРА И РУНГЕ–КУТТЫ РЕШЕНИЯ НАЧАЛЬНЫХ ЗАДАЧ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Ставится задача получения приближенных решений обыкновенных дифференциальных уравнений первого порядка с заданными начальными условиями. Уделяется внимание различным способам вывода численного метода Эйлера, являющегося наиболее простым частным случаем нескольких групп численных процессов разной идеологии. Рассматриваются непосредственные модификации методов Эйлера; одна из таких модификаций, называемая исправленным методом Эйлера, приводит к семейству методов Рунге–Кутты второго порядка. Записывается общий вид формул Рунге–Кутты произвольного порядка, дается геометрическая интерпретация классического метода Рунге–Кутты четвертого порядка, обсуждаются вопросы пошагового контроля точности при реализации методов, приводится алгоритм Кутты–Мерсона, решающий проблему автоматического выбора расчетного шага при численном интегрировании поставленной начальной задачи с заданной точностью.

14.1. ПОСТАНОВКА ЗАДАЧИ. КЛАССИФИКАЦИЯ ПРИБЛИЖЕННЫХ МЕТОДОВ. МЕТОД ПОСЛЕДОВАТЕЛЬНЫХ ПРИБЛИЖЕНИЙ

Будем рассматривать обыкновенное дифференциальное уравнение (ОДУ) первого порядка

$$y' = f(x, y), \quad x \in [x_0, b] \qquad (14.1)$$

с начальным условием

$$y(x_0) = y_0, \qquad (14.2)$$

где $f(x, y)$ — некоторая заданная, в общем случае, нелинейная функция двух переменных. Будем считать, что для данной задачи (14.1)–(14.2), называемой ***начальной задачей*** или ***задачей Коши***, выполняются требования, обеспечивающие существование и единственность на отрезке $[x_0, b]$ ее решения $y = y(x)$ (такие требования можно найти в любом курсе дифференциальных уравнений или в соответствующем разделе курса высшей математики, см., например [109, 164, 166]). Более того, не оговаривая это отдельно, будем предполагать, что искомое решение облада-

ет той или иной степенью гладкости, необходимой для построения и «законного» применения того или иного метода.

Несмотря на внешнюю простоту уравнения (14.1), решить его аналитически, т.е. найти общее решение $y = y(x, C)$ с тем, чтобы затем выделить из него интегральную кривую $y = y(x)$, проходящую через заданную точку $(x_0; y_0)$, удается лишь для некоторых специальных типов таких уравнений, описание которых также можно обнаружить, например, в упомянутых литературных источниках. Поэтому, как и в родственной для (14.1)–(14.2) задаче вычисления интегралов, приходится делать ставку на приближенные способы решения начальных задач для ОДУ, которые можно разделить на три группы:

1) *приближенно-аналитические методы*;
2) *графические или машинно-графические методы*;
3) *численные методы*.

К методам первой группы относят такие, которые позволяют находить приближение решения $y(x)$ сразу в виде некоторой «хорошей» функции $\varphi(x)$. Например, широко известен ***метод степенных рядов***, в одну из реализаций которого заложено представление искомой функции $y(x)$ отрезком ряда Тейлора, где тейлоровские коэффициенты, содержащие производные высших порядков, находят последовательным дифференцированием самого уравнения (14.1) [62, 100 и др.]. Другим представителем этой группы методов является метод последовательных приближений, суть которого приведена чуть ниже.

Название ***графические методы*** говорит о приближенном представлении искомого решения $y(x)$ на промежутке $[x_0, b]$ в виде графика, который можно строить по тем или иным правилам, связанным с графическим толкованием данной задачи. Физическая или, возможно, точнее будет сказать, электротехническая интерпретация начальных задач для определенных видов уравнений лежит в основе машинно-графических методов приближенного решения. Реализуя на физико-техническом уровне заданные электрические процессы, на экране осциллографа наблюдают поведение решений дифференциальных уравнений, описывающих эти процессы. Изменение параметров уравнений приводит к адекватному изменению поведения решений, что положено в основу специализированных ***аналоговых вычислительных машин*** (АВМ).

Наконец, наиболее значимыми в настоящее время, характеризуемое бурным развитием и проникновением во все сферы человеческой деятельности цифровой вычислительной техники, являются численные методы решения дифференциальных уравнений, предполагающие получение числовой таблицы

приближенных значений y_i искомого решения $y(x)$ на некоторой сетке $x_i \in [x_0, b]$ значений аргумента x. Этим способам и будет посвящено дальнейшее изложение. Что делать с получаемыми численными значениями решения, зависит от прикладной постановки задачи. Если речь идет о нахождении только значения $y(b)$, тогда точка b включается как конечная в систему расчетных точек x_i, и все приближенные значения $y_i \approx y(x_i)$, кроме последнего, участвуют лишь как промежуточные, т.е. не требуют ни запоминания, ни обработки. Если же нужно иметь приближенное решение $y(x)$ в любой точке x, то для этого к получаемой числовой таблице значений y_i можно применить какой-либо из способов аппроксимации табличных функций, рассмотренных ранее, например, интерполяцию (гл.1) или сплайн-интерполяцию (гл.4). Возможны и другие использования численных данных о решении.

Коснемся одного приближенно-аналитического способа решения начальной задачи (14.1)–(14.2), в котором искомое решение $y = y(x)$ в некоторой правой окрестности точки x_0 является пределом последовательности получаемых определенным образом функций $y_n(x)$.

Проинтегрируем левую и правую части уравнения (14.1) в границах от x_0 до x:

$$\int_{x_0}^{x} y'(t)dt = \int_{x_0}^{x} f(t, y(t))\,dt.$$

Отсюда, с учетом того, что одной из первообразных для $y'(x)$ служит $y(x)$, получаем

$$y(x) - y(x_0) = \int_{x_0}^{x} f(t, y(t))\,dt$$

или, с использованием начального условия (14.2),

$$y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))\,dt. \qquad (14.3)$$

Таким образом, данное дифференциальное уравнение (14.1) с начальным условием (14.2) преобразовалось в ***интегральное уравнение*** (неизвестная функция здесь входит под знак интеграла; несколько более подробно об этом см. гл.18).

Полученное интегральное уравнение (14.3) имеет вид задачи о неподвижной точке

$$y = \phi(y)$$

для оператора $\phi(y) := y_0 + \int\limits_{x_0}^{x} f(t, y(t))dt$. Формально к этой задаче можно применить метод простых итераций

$$y_{n+1} = \phi(y_n), \quad n=0, 1, 2, \ldots, \tag{14.4}$$

достаточно обстоятельно рассматривавшийся в гл. 3, 5, 6 применительно к скалярным уравнениям, а также к системам линейных и нелинейных алгебраических и трансцендентных уравнений. Беря в качестве начальной функции $y_0(x)$ заданную в (14.2) постоянную y_0, по формуле (14.4) при $n=0$ находим первое приближение

$$y_1(x) = y_0 + \int\limits_{x_0}^{x} f(t, y_0)dt.$$

Его подстановка в (14.4) при $n=1$ дает второе приближение

$$y_2(x) = y_0 + \int\limits_{x_0}^{x} f(t, y_1(t))dt,$$

и т.д. Таким образом, этот приближенно-аналитический метод, называемый ***методом последовательных приближений*** или ***методом Пикара****), определяется формулой

$$y_{n+1}(x) = y_0 + \int\limits_{x_0}^{x} f(t, y_n(t))dt, \tag{14.5}$$

где $n = 0, 1, 2, \ldots$ и $y_0(x) \equiv y_0$.

Хорошо изучена сходимость метода (14.5). В частности, не вдаваясь в подробности, констатируем один качественный результат [62]:

если в некоторой односвязной области G, содержащей точку $(x_0; y_0)$,

$$|f(x, y)| \le C, \quad |f'_y(x, y)| \le C_1,$$

то найдется такая постоянная $h > 0$, что на отрезке $[x_0, x_0 + h]$ последовательность функций $y_n(x)$, определяемая методом (14.5), равномерно сходится к решению $y(x)$ задачи Коши (14.1)–(14.2) и справедлива оценка погрешности

$$|y(x) - y_n(x)| \le C \cdot C_1^n \cdot \frac{(x - x_0)^{n+1}}{(n+1)!}.$$

---

*) Пика́р Шарль Эмиль (1856–1941) — французский математик.

Отметим две характеристики метода последовательных приближений Пикара, которые можно отнести к негативным. Во-первых, в силу известных проблем с эффективным нахождением первообразных, в чистом виде метод (14.5) редко реализуем. Во-вторых, как видно из вышеприведенного утверждения, этот метод следует считать локальным, пригодным для приближения решения в малой правой окрестности начальной точки. Большее значение метод Пикара имеет для доказательства существования и единственности решения задачи Коши, нежели для его практического нахождения.

14.2. МЕТОД ЭЙЛЕРА — РАЗНЫЕ ПОДХОДЫ К ПОСТРОЕНИЮ

Учитывая ключевую позицию, которую занимает метод Эйлера в теории численных методов ОДУ, рассмотрим несколько способов его вывода. При этом будем считать, что вычисления проводятся с ***расчетным шагом*** $h = \dfrac{b - x_0}{n}$, ***расчетными точками (узлами)*** служат точки $x_i = x_0 + ih \quad (i = 0, 1, \ldots, n)$ промежутка $[x_0, b]$ и целью является построение таблицы

x	x_0	x_1	$\ldots$	$x_n = b$
y	y_0	y_1	$\ldots$	$y_n \approx y(b)$

приближенных значений y_i решения $y = y(x)$ задачи (14.1)–(14.2) в расчетных точках x_i.

Геометрический способ. Пользуясь тем, что в точке x_0 известно и значение решения $y(x_0) = y_0$ (согласно (14.2)), и значение его производной $y'(x_0) = f(x_0, y_0)$ (согласно (14.1)), можно записать уравнение касательной к графику искомой функции $y = y(x)$ в точке $(x_0; y_0)$:

$$y = y_0 + f(x_0, y_0)(x - x_0). \tag{14.6}$$

При достаточно малом шаге h ордината

$$y_1 = y_0 + hf(x_0, y_0) \tag{14.7}$$

этой касательной, полученная подстановкой в правую часть (14.6) значения $x_1 = x_0 + h$, по непрерывности должна мало отличаться от ординаты $y(x_1)$ решения $y(x)$ задачи (14.1)–(14.2). Следовательно, точка (x_1, y_1) пересечения касательной (14.6) с

прямой $x = x_1$ может быть приближенно принята за новую начальную точку. Через эту точку снова проведем прямую

$$y = y_1 + f(x_1, y_1)(x - x_1),$$

которая уже приближенно отражает поведение касательной к $y = y(x)$ в точке $(x_1; y(x_1))$. Подставляя сюда $x = x_2 (= x_1 + h)$, иначе, пересекая эту «касательную» прямой $x = x_2$, получим приближение значения $y(x_2)$ значением

$$y_2 = y_1 + hf(x_1, y_1),$$

и т.д. В итоге этого процесса, определяемого формулой

$$y_{i+1} = y_i + hf(x_i, y_i), \quad i = 0, 1, \ldots, n \qquad (14.8)$$

и называемого ***методом Эйлера***, график решения $y = y(x)$ данной задачи Коши (14.1)–(14.2) приближенно представляется ломаной, составленной из отрезков приближенных касательных (рис. 14.1), откуда происходит другое название метода (14.8) — ***метод ломаных***.

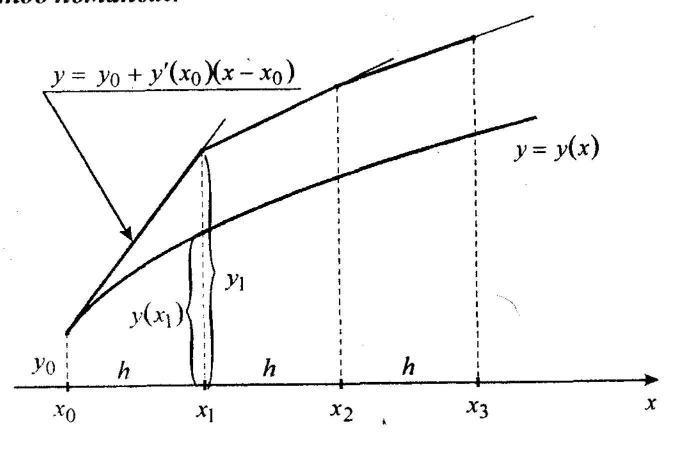


Рис. 14.1. Геометрическая интерпретация метода Эйлера

Применение формулы Тейлора. Описываемый здесь способ вывода метода Эйлера тесно связан с предыдущим. Линеаризуя решение в окрестности начальной точки по формуле Тейлора, имеем

$$y(x) = y(x_0) + y'(x_0)(x - x_0) + \frac{y''(\xi)}{2}(x - x_0)^2.$$

Отсюда при $x = x_1$ получаем

$$y(x_1) = y_0 + hf(x_0, y_0) + \frac{y''(\xi_1)}{2}h^2. \qquad (14.9)$$

Точное равенство (14.9), переписанное в виде

$$y(x_1) = y_1 + r_1(h),$$

говорит о том, что здесь мы имеем одновременно как саму формулу Эйлера для вычисления значения $y_1 \approx y(x_1)$ (сравните с формулой (14.7)), так и ее остаточный член

$$r_1(h) = \frac{y''(\xi_1)}{2}h^2, \qquad (14.10)$$

где ξ_1 — некоторая точка интервала (x_0, x_1).

Остаточный член (14.10) характеризует ***локальную*** (или, иначе, ***шаговую***) ***ошибку*** метода Эйлера, т.е. ошибку, совершаемую на одном шаге. Очевидно, что от шага к шагу, т.е. при многократном применении формулы (14.8), возможно наложение ошибок*). За n шагов, т.е. в точке b, образуется ***глобальная ошибка***; изучение такой ошибки будет проведено позже (см. § 16.2). Сейчас же анонсируем один важный хотя бы для терминологии факт: *порядок глобальной ошибки (относительно шага h) на единицу ниже, чем порядок локальной ошибки, а порядком глобальной ошибки и определяется порядок соответствующего численного процесса решения задачи Коши.* Таким образом, локальная ошибка метода Эйлера, согласно (14.10), есть $O(h^2)$, глобальная — $O(h)$, т.е. *метод Эйлера относится к методам первого порядка.*

Разностный способ. Рассматривая уравнение (14.1) в точке x_0, с учетом (14.2) имеем равенство

$$y'(x_0) = f(x_0, y_0).$$

Применяя к его левой части аппроксимацию производной пра-

*) Заметим, что в отличие от метода последовательных приближений (14.5), являющегося итерационным по сути, все рассматриваемые численные процессы решения задачи Коши для ОДУ, и в частности, метод Эйлера (14.8), являются итерационными лишь по форме. На самом деле, это *шаговые* методы, в которых на каждом шаге выполняются однотипные действия; характерного для итерационных методов уточнения решения здесь не происходит.

вым разностным отношением первого порядка точности (см. (6.15) при $i=0$)

$$y'(x_0)=\frac{y(x_1)-y(x_0)}{h}-\frac{y''(\xi_1)}{2}h,$$

получаем

$$\frac{y(x_1)-y(x_0)}{h}=f(x_0,y_0)+\frac{y''(\xi_1)}{2}h,$$

что идентично равенству (14.9), поставляющему формулу для вычисления y_1 вида (14.7) и локальный остаточный член (14.10). Ясно, что для получения общей расчетной формулы (14.8) можно было сразу применить аппроксимацию $y'(x_i)$ по формуле (6.16) в равенстве

$$y'(x_i)=f(x_i,y(x_i)), \qquad (14.11)$$

заменив неизвестное точное значение $y(x_i)$ известным приближенным значением y_i.

Заметим, что *порядок получающегося таким способом метода численного интегрирования дифференциальной задачи* (14.1)–(14.2) *совпадает с порядком аппроксимации производной в левой части уравнения* (14.1).

Квадратурный способ. Как было показано в § 14.1, начальную задачу для ОДУ (14.1)–(14.2) можно заменить эквивалентным интегральным уравнением (14.3). При $x=x_1$ из него получится равенство

$$y(x_1)=y_0+\int_{x_0}^{x_1}f(x,y(x))dx. \qquad (14.12)$$

Применение к интегралу в правой части равенства (14.12) простейшей (одноточечной) формулы левых прямоугольников (5.6) дает приближенную формулу

$$y(x_1)\approx y_0+f(x_0,y(x_0))(x_1-x_0),$$

правая часть которой, очевидно, совпадает с выражением (14.7) для подсчета значения y_1. В общем случае расчетная формула (14.8) метода Эйлера получается численным интегрированием посредством простейшей формулы левых прямоугольников в равенстве

$$y(x_{i+1})-y(x_i)=\int_{x_i}^{x_{i+1}}f(x,y(x))dx \qquad (14.13)$$

в предположении, что на каждом i-м шаге в роли начальной

точки (x_0,y_0) выступает точка (x_i,y_i). Зная точность используемой в (14.13) квадратурной формулы, легко прийти к тому же выражению локальной ошибки метода Эйлера, что и при других способах его построения.

Существуют и другие подходы к выводу метода Эйлера. В частности, он будет возникать далее как частный случай некоторых семейств численных методов решения задачи (14.1)–(14.2).

14.3. НЕСКОЛЬКО ПРОСТЫХ МОДИФИКАЦИЙ МЕТОДА ЭЙЛЕРА

Разовьем последний из подходов к построению метода Эйлера. Очевидно, применение к интегральному равенству (14.13) других простейших квадратурных формул будет порождать новые методы численного интегрирования задачи Коши (14.1)–(14.2).

Так, если в (14.13) использовать простейшую квадратурную формулу правых прямоугольников (5.7), придем к методу

$$y_{i+1}=y_i+hf(x_{i+1},y_{i+1}),\quad i=0,1,\ldots,n. \qquad (14.14)$$

Этот метод называют ***неявным*** (или ***обратным***) ***методом Эйлера***, поскольку для вычисления неизвестного значения $y_{i+1}\approx y(x_{i+1})$ по известному значению $y_i\approx y(x_i)$ требуется решать уравнение, в общем случае нелинейное. Имеет ли свою сферу применения подобный метод, порядок которого такой же, как и у явного метода Эйлера (14.8) (первый*)), и один шаг вычислений по которому столь трудоемок? Положительный ответ на этот вопрос будет дан в гл. 16.

Применение к интегралу в (14.13) простейшей квадратурной формулы трапеций (5.26) приводит тоже к неявному методу

$$y_{i+1}=y_i+\frac{h}{2}[f(x_i,y_i)+f(x_{i+1},y_{i+1})],\quad i=0,1,\ldots,n, \qquad (14.15)$$

который будем называть ***методом трапеций***. Квадратурная формула трапеций, как известно из гл.5, на порядок точнее формул левых и правых прямоугольников. Отсюда вытекает более высокий (на единицу) порядок точности метода трапеций (14.15) по сравнению с явным и с неявным методами Эйлера (14.8) и (14.14), т.е. *метод трапеций* (14.15) — *это метод второго порядка* (впослед-

*) Очевидно, локальный остаточный член метода (14.14) лишь знаком будет отличаться от локального остаточного члена (14.10) метода (14.8).

ствии, в § 15.4, к этому выводу придем из других соображений).

Некоторый интерес представляет совместное применение явного метода Эйлера и неявного метода трапеций.

По форме равенство (14.15) представляет собой скалярную задачу о неподвижной точке относительно неизвестного значения y_{i+1}. Поэтому, если в правую часть (14.15) подставить хорошее начальное приближение y_{i+1}^0, подсчитываемое по формуле (14.14), то тогда само это равенство можно считать шагом метода простых итераций для уточнения этого значения (см. гл.6). Таким образом, получаем гибридный метод

$$y_{i+1} = y_i + \frac{h}{2}[f(x_i, y_i) + f(x_i + h, y_i + hf(x_i, y_i))], \quad i = 0, 1, ..., n, \tag{14.16}$$

который называют ***методом Хойна*** [6, 60] (или *Хьюна* [138]).

Ясно, что можно достичь большей точности, если, исходя из того же начального приближения

$$y_{i+1}^0 = y_i + hf(x_i, y_i),$$

сделать не одну, а несколько итераций по методу трапеций:

$$y_{i+1}^{(k)} = y_i + \frac{h}{2}\left[f(x_i, y_i) + f\left(x_{i+1}, y_{i+1}^{(k-1)}\right)\right], \quad k = 1, 2, \dots \tag{14.17}$$

Такой вариант совместного применения метода Эйлера и метода трапеций называют ***усовершенствованным методом Эйлера-Коши*** [84] ***с итерационной обработкой****). Делать много итераций по формуле (14.17) не рекомендуется (обычно их выполняют не более трех-четырех). Совпадение определенного числа разрядов в полученных числах $y_{i+1}^{(k)}$ и $y_{i+1}^{(k-1)}$ говорит о точности, с которой решено методом простых итераций уравнение (14.15) относительно y_{i+1}, а вовсе не о том, что с такой точностью найдено значение $y(x_{i+1})$.

Чтобы получить следующую модификацию метода Эйлера, проинтегрируем уравнение (14.1) по отрезку $[x_{i-1}, x_{i+1}]$. Имеем

$$\int_{x_{i-1}}^{x_{i+1}} y'(x)dx = \int_{x_{i-1}}^{x_{i+1}} f(x, y(x))dx,$$

откуда следует равенство

$$y(x_{i+1}) = y(x_{i-1}) + \int_{x_{i-1}}^{x_{i+1}} f(x, y(x))dx. \tag{14.18}$$

*) Иначе, ***методом Эйлера с пересчетом*** [9].

Применяя к последнему интегралу одноточечную квадратурную формулу средних прямоугольников (12.10) и заменяя значения $y(x_{i-1})$ и $y(x_i)$ известными приближенными значениями y_{i-1} и y_i соответственно, из (14.18) выводим формулу для подсчета приближенного значения $y(x_{i+1})$

$$y_{i+1} = y_{i-1} + 2hf(x_i, y_i), \quad i = 1, 2, ..., n-1, \tag{14.19}$$

которую будем называть ***уточненным методом Эйлера*** [58].

Как известно, квадратурная формула прямоугольников (средней точки) имеет тот же порядок точности, что и квадратурная формула трапеций, так что *уточненный метод Эйлера* (14.19) тоже *является методом второго порядка**). Подтверждением этого факта может служить вывод метода (14.19) на разностной основе. Применив к равенству (14.11) формулу симметричной аппроксимации $y'(x_i)$ второго порядка точности, получим

$$\frac{y(x_{i+1}) - y(x_{i-1})}{2h} \approx f(x_i, y(x_i)),$$

откуда после приближенной замены $y(x_{i-1}) \approx y_{i-1}$, $y(x_i) \approx y_i$, $y(x_{i+1}) \approx y_{i+1}$ следует (14.19).

Обратим внимание на одно принципиальное отличие метода (14.19) от всех других рассмотренных до этого момента методов: метод (14.19) является ***двухшаговым***. Здесь для вычисления значения y_{i+1} привлекаются два предыдущих значения y_i и y_{i-1}. Двухшаговость накладывает определенные ограничения, по крайней мере, на начало численного процесса: значение $y_1 \approx y(x_1)$ не может быть найдено непосредственно этим методом с тем же шагом h. Поэтому недостающую вторую начальную для процесса (14.19) точку приходится получать другим путем, например, явным методом Эйлера, а чтобы не сделать сразу большой ошибки, применяя на старте метод более низкого порядка точности, рекомендуется осуществлять постепенное вхождение в процесс (14.19). Так, «разгон» можно выполнить по формулам

$$y_{\frac{1}{2}} = y_0 + \frac{h}{2}f(x_0, y_0), \quad y_1 = y_0 + hf\left(x_0 + \frac{h}{2}, y_{\frac{1}{2}}\right), \tag{14.20}$$

*) Другие его названия: ***метод Нистрёма*** (Нюстрема) ***второго порядка*** [6, 185], ***метод Милна второго порядка*** [198].

а далее уже переключаться на счет по формуле (14.19).

Пример 14.1. Рассмотрим простое линейное уравнение

$$y' = 2x - 3y$$

с начальным условием $y(0) = 1$. На этой задаче легко проследить за вычислениями, реализующими различные выведенные выше методы. Знание ее точного решения $y(x) = -\frac{2}{9} + \frac{2}{3}x + \frac{11}{9}e^{-3x}$ позволяет провести сравнение результатов приближенных вычислений по разным формулам с истинным решением и проверить, насколько соответствуют представления о точности тех или иных методов тому, что наблюдается в данном, наверное, далеко не самом типичном частном случае.

Сначала сделаем несколько последовательных приближений по методу Пикара. Его итерационная формула (14.5) для данной начальной задачи имеет вид

$$y_{n+1}(x) = 1 + x^2 - 3\int_0^x y_n(t)dt .$$

Подставляя сюда $y_0 = 1$, при $n = 0, 1, 2$ последовательно получаем:

$$y_1(x) = 1 + x^2 - 3\int_0^x dt = 1 - 3x + x^2,$$

$$y_2(x) = 1 + x^2 - 3\int_0^x \left(1 - 3t + t^2\right)dt = 1 - 3x + \frac{11}{2}x^2 - x^3,$$

$$y_3(x) = 1 + x^2 - 3\int_0^x \left(1 - 3t + \frac{11}{2}t^2 - t^3\right)dt = 1 - 3x + \frac{11}{2}x^2 - \frac{11}{2}x^3 + \frac{3}{4}x^4.$$

Эти результаты удобно сравнить с точным решением, если в последнем разложить в ряд по степеням x фигурирующую там функцию e^{-3x}. Тогда получим представление решения в виде ряда

$$y(x) = 1 - 3x + \frac{11}{2}x^2 - \frac{11}{2}x^3 + \frac{33}{8}x^4 - \ldots,$$

с которым, как видим, хорошо согласуются приближения y_1, y_2, y_3, определяемые методом Пикара.

Теперь проведем подсчет приближенных значений решения $y(x)$ данной задачи в точке $x = 0.2$ численным методом Эйлера и его модификациями, принимая $h = 0.1$ (т.е. за два шага). Результаты этих вычислений и фактические ошибки, найденные сравнением с точным значением $y(0.2) = 0.581881\ldots$, отражены в следующей таблице.

Метод	$y_1 \approx y(0.1)$	$y_2 \approx y(0.2)$	$y(0.2) - y_2$
Эйлера (14.8)	0.7	0.51	≈ 0.07
Неявный Эйлера (14.14)	≈ 0.7846	≈ 0.6343	≈ -0.05
Трапеций (14.15)	≈ 0.7478	≈ 0.5788	≈ 0.003
Хойна (14.16)	0.755	≈ 0.5895	≈ -0.008
Уточненный Эйлера (14.19)–(14.20)	0.755	0.587	≈ -0.005

Последний столбец в этой таблице со всей очевидностью показывает бо́льшую точность методов второго порядка (см. три последних строки).

14.4. ИСПРАВЛЕННЫЙ МЕТОД ЭЙЛЕРА

Пусть найдено приближенное значение $y_i \approx y(x_i)$ решения $y = y(x)$ задачи (14.1)–(14.2) и требуется вычислить $y_{i+1} \approx y(x_{i+1})$, где $x_{i+1} = x_i + h$. Запишем разложение решения по формуле Тейлора p-го порядка, принимая за базовую точку x_i (т.е. по степеням $x - x_i$) и положим в этом разложении $x = x_{i+1}$. Имеем

$$y(x_{i+1}) = y(x_i) + hy'(x_i) + \frac{1}{2!}h^2 y''(x_i) + \ldots$$

$$\ldots + \frac{1}{p!}h^p y^{(p)}(x_i) + O\left(h^{p+1}\right). \quad (14.21)$$

Если ограничится двумя слагаемыми в правой части разложения (14.21), то, согласно показанному в § 14.2, получим обычный метод Эйлера (14.8). Посмотрим, что дает учитывание третьего слагаемого.

При $p = 2$ из (14.21) следует равенство

$$y(x_{i+1}) = y(x_i) + hy'(x_i) + \frac{h^2}{2}y''(x_i) + O\left(h^3\right). \quad (14.22)$$

Значение первой производной в точке x_i, в силу связи (14.1), приближенно известно:

$$y'(x_i) = f(x_i,\ y(x_i)) \approx f(x_i,\ y_i). \quad (14.23)$$

Дифференцируя (14.1), по формуле полной производной

$$y''(x) = f'_x(x, y) + f'_y(x, y)y'$$

находим приближенное значение второй производной:

$$y''(x_i) = f'_x(x_i, y(x_i)) + f'_y(x_i, y(x_i))f(x_i, y(x_i)) \approx$$

$$\approx f'_x(x_i, y_i) + f'_y(x_i, y_i)f(x_i, y_i). \qquad (14.24)$$

Подставляя приближенные выражения $y(x_i)$, $y'(x_i)$ и $y''(x_i)$ в равенство (14.22), получаем следующую формулу для вычисления $y_{i+1} \approx y(x_{i+1})$ при $i = 0, 1, \ldots, n$:

$$y_{i+1} = y_i + h\left[f(x_i, y_i) + \frac{h}{2}\left(f'_x(x_i, y_i) + f'_y(x_i, y_i)f(x_i, y_i)\right)\right]. \qquad (14.25)$$

Определяемый ею метод будем называть ***исправленным методом Эйлера***.

Так как при $i = 0$ формулы (14.23) и (14.24) точны, а $y_0 = y(x_0)$, согласно начальному условию (14.2), то на первом шаге вычислений по формуле (14.25) будет совершаться ошибка, связанная только с усечением ряда Тейлора. Следовательно, локальная ошибка или, иначе, ***шаговая погрешность*** метода (14.25) составляет величину $O(h^3)$, а это означает, что *исправленный метод Эйлера относится к методам второго порядка.*

14.5. О СЕМЕЙСТВЕ МЕТОДОВ РУНГЕ–КУТТЫ. МЕТОДЫ ВТОРОГО ПОРЯДКА

Недостатком исправленного метода Эйлера (14.25) и других методов более высоких порядков, основанных на пошаговом представлении решения $y(x)$ задачи (14.1)–(14.2) по формуле Тейлора и последовательном дифференцировании уравнения (14.1) для получения тейлоровых коэффициентов, является необходимость вычисления на каждом шаге частных производных функции $f(x, y)$.

Идея построения явных ***методов Рунге–Кутты****) p-го порядка заключается в получении приближений к значениям $f(x_{i+1})$ по формуле вида

$$y_{i+1} = y_i + h\varphi(x_i, y_i, h), \qquad (14.26)$$

где $\varphi(x, y, h)$ — некоторая функция, приближающая отрезок ряда Тейлора (7.21) до p-го порядка и не содержащая частных производных функции $f(x, y)$.

Так, полагая в (7.26) $\varphi(x, y, h) \equiv f(x, y)$, приходим к методу Эйлера (14.8), т.е. метод Эйлера можно считать простейшим примером методов Рунге–Кутты, соответствующим случаю $p = 1$.

Для построения методов Рунге–Кутты порядка, выше первого, функцию $\varphi(x, y, h)$ берут многопараметрической, и подбирают ее параметры сравнением выражения (14.26) с многочленом Тейлора для $y(x)$ соответствующей желаемому порядку степени.

Рассмотрим случай $p = 2$. Возьмем функцию φ в (14.26) следующей структуры:

$$\varphi(x, y, h) := c_1 f(x, y) + c_2 f(x + ah, y + bhf(x, y)).$$

Ее параметры c_1, c_2, a и b будем подбирать так, чтобы записанная, согласно (14.26), формула

$$y_{i+1} = y_i + h[c_1 f(x_i, y_i) + c_2 f(x_i + ah, y_i + bhf(x_i, y_i))] \qquad (14.27)$$

определяла метод второго порядка, т.е. чтобы максимальная локальная ошибка составляла величину $O(h^3)$.

Разложим функцию двух переменных $f(x + ah, y + bhf(x, y))$ по формуле Тейлора, ограничиваясь линейными членами:

$$f(x + ah, y + bhf(x, y)) = f(x, y) + f'_x(x, y)ah +$$

$$+ f'_y(x, y)bhf(x, y) + O(h^2).$$

Ее подстановка в (14.27) дает

$$y_{i+1} = y_i + h[(c_1 + c_2)f(x_i, y_i) + h(c_2 a f'_x(x_i, y_i) +$$

$$+ c_2 b f'_y(x_i, y_i)f(x_i, y_i))] + O(h^3). \qquad (14.28)$$

Сравнение последнего выражения с тейлоровским квадратичным представлением решения $y(x)$ (14.22) с точностью до $O(h^3)$ равносильно сравнению его с выражением y_{i+1} по формуле (14.25), т.е. с исправленным методом Эйлера. Очевидно, чтобы (14.28) и (14.25) совпадали с точностью $O(h^3)$, от параметров нужно

---

*) 1) Кутта Мартин Вильгельм (1867–1944) — немецкий физик и математик.

2) Ранее типичным было написание «метод Рунге–Кутта»; в последнее время все чаще говорят и пишут «метод Рунге–Кутты».

3) О неявных методах Рунге–Кутты можно прочитать, например, в [188].

потребовать выполнение следующей совокупности условий:

$$\begin{cases} c_1 + c_2 &= 1, \\ c_2 a &= 0.5, \\ c_2 b &= 0.5. \end{cases} \tag{14.29}$$

Полученная система условий содержит три уравнения относительно четырех параметров метода. Это говорит о наличии одного свободного параметра. Положим $c_2 = \alpha$ ($\neq 0$). Тогда из (14.29) имеем:

$$c_1 = 1 - \alpha, \quad a = \frac{1}{2\alpha}, \quad b = \frac{1}{2\alpha}.$$

В результате подстановки этих значений параметров в формулу (14.27) приходим к ***однопараметрическому семейству методов Рунге–Кутты второго порядка***.

$$y_{i+1} = y_i + h\left[(1-\alpha)f(x_i, y_i) + \alpha f\left(x_i + \frac{h}{2\alpha},\ y_i + \frac{h}{2\alpha}f(x_i, y_i)\right)\right]. \tag{14.30}$$

Выделим из семейства методов (14.30) два наиболее простых и естественных частных случая:

при $\alpha = \frac{1}{2}$ получаем формулу

$$y_{i+1} = y_i + \frac{h}{2}[f(x_i, y_i) + f(x_i + h,\ y_i + hf(x_i, y_i))],$$

в котором узнаём ***метод Хойна*** (14.16), полученный ранее из других соображений;

при $\alpha = 1$ из (14.30) выводим новый простой метод

$$y_{i+1} = y_i + hf\left(x_i + \frac{h}{2},\ y_i + \frac{h}{2}f(x_i, y_i)\right), \tag{14.31}$$

который назовем ***методом средней точки***.

14.6. МЕТОДЫ РУНГЕ–КУТТЫ ПРОИЗВОЛЬНОГО И ЧЕТВЕРТОГО ПОРЯДКОВ

Любой метод из семейства методов Рунге–Кутты второго порядка (14.30) реализуют по следующей схеме. На каждом шаге, т.е. при каждом $i = 0, 1, 2, \ldots$, вычисляют значения функции

$$\eta_1^i := f(x_i, y_i), \quad \eta_2^i := f\left(x_i + \frac{h}{2\alpha}, y_i + \frac{h}{2\alpha}\eta_1^i\right),$$

а затем находят ***шаговую поправку***

$$\Delta y_i := h\left[(1-\alpha)\eta_1^i + \alpha\eta_2^i\right],$$

прибавление которой к результату предыдущего шага дает приближенное значение решения $y(x)$ в точке $x_{i+1} = x_i + h$:

$$y_{i+1} = y_i + \Delta y_i.$$

Метод такой структуры называют ***двухэтапным*** по количеству вычислений значений функции — правой части уравнения (14.1) — на одном шаге.

Анализ устройства методов Рунге–Кутты второго порядка позволяет представить, в какой форме следует конструировать явный метод Рунге–Кутты произвольного порядка. По аналогии с предыдущим для семейства методов Рунге–Кутты p-го порядка используется запись, состоящая из следующей совокупности формул:

$$\begin{cases} \eta_1^i = f(x_i, y_i), \\ \eta_k^i = f\left(x_i + a_k h, \quad y_i + h\sum\limits_{j=1}^{k-1} b_{kj}\eta_j^i\right), \\ y_{i+1} = y_i + h\sum\limits_{k=1}^{p} c_k\eta_k^i, \end{cases} \tag{14.32}$$

где $k = 2, 3, \ldots, p$ (для p-этапного метода). Многочисленные параметры c_k, a_k, b_{kj}, фигурирующие в формулах (14.32), подбираются так, чтобы получаемое методом (14.32) значение y_{i+1} совпадало со значением разложения $y(x_{i+1})$ по формуле Тейлора с погрешностью $O(h^{p+1})$ (без учета погрешностей, совершаемых на предыдущих шагах).

Наиболее употребительным частным случаем семейства методов (14.32) является следующий ***метод Рунге–Кутты четвертого порядка****), относящийся к ***четырехэтапным*** и

*) Как правило, когда говорят «метод Рунге–Кутты», не сообщая о нем никаких дополнительных сведений, то под этим подразумевают именно данный классический метод четвертого порядка.

имеющий вид:

$$\begin{cases} \eta_1^i = f(x_i, y_i), \\ \eta_2^i = f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2}\eta_1^i\right), \\ \eta_3^i = f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2}\eta_2^i\right), \\ \eta_4^i = f\left(x_i + h, y_i + h\eta_3^i\right), \\ \Delta y_i = \frac{h}{6}\left(\eta_1^i + 2\eta_2^i + 2\eta_3^i + \eta_4^i\right), \\ y_{i+1} = y_i + \Delta y_i. \end{cases} \tag{14.33}$$

Не пытаясь воспроизвести выкладки, приводящие от общей записи семейства (14.32) при $p = 4$ к конкретному методу (14.33), дадим геометрическое толкование последнего.

Обратив внимание на то, что шаговая поправка Δy_i есть средневзвешенная величина поправок $h\eta_1^i$, $h\eta_2^i$, $h\eta_3^i$, $h\eta_4^i$ каждого этапа (с весовыми коэффициентами $\frac{1}{6}, \frac{2}{6}, \frac{2}{6}, \frac{1}{6}$ соответственно), проанализируем, как получаются эти поправки этапов. На первом этапе создается приращение $h\eta_1^i = hf(x_i, y_i)$ $(= hy'(x_i))$, соответствующее шаговой поправке Эйлера, — это очевидно. На рис 14.2 ему отвечает отрезок BC вертикали $x = x_{i+1}$ (точка B получена ортогональным проектированием точки A на эту вертикаль).

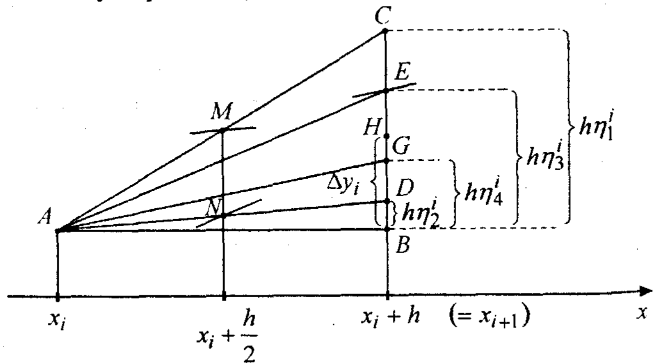


Рис. 14.2. Геометрическая иллюстрация одного шага метода Рунге–Кутты четвертого порядка

Так как точка M, благодаря свойству средней линии треугольника (см. ΔABC), имеет ординату $y_i + \frac{h}{2}\eta_1^i$, то η_2^i определяет значение $f(M)$, служащее (согласно связи $y' = f(x, y)$ и геометрическому смыслу производной) тангенсом угла A в новом треугольнике с противолежащим этому углу катетом $h\eta_2^i = BD$. Далее, аналогично, подсчитав $f(N) = f\left(x_i + \frac{h}{2}, y_i + \frac{h}{2}\eta_2^i\right) = \eta_3^i$, на вертикали $x = x_{i+1}$ откладываем следующую промежуточную (этапную) поправку $h\eta_3^i = BE$. Вычислив величину $f(E) = f\left(x_i + h, y_i + h\eta_3^i\right)$, являющуюся значением тангенса угла A во вновь получаемом ΔABG, имеем поправку $h\eta_4^i = BG$ последнего этапа. Итоговая шаговая поправка $\Delta y_i = BH$ есть продукт усреднения с указанными коэффициентами четырех этапных поправок — длин отрезков BC, BD, BE и BG. Точка H будет стартовой для следующего, $i+1$-го, шага метода (14.33).

Заметим, что если первый этап, как уже упоминалось, соответствует применению явного метода Эйлера, то четвертый — неявного, а второй и третий — уточненного методов Эйлера. Последний имеет более высокий порядок точности, отсюда и больший вес отвечающих ему значений этапных поправок.

14.7. ПОШАГОВЫЙ КОНТРОЛЬ ТОЧНОСТИ. МЕТОД КУТТЫ–МЕРСОНА

Нетрудно понять, что выведение надежных и, в то же время, простых и эффективных оценок погрешности, гарантирующих получение таблицы значений решения $y = y(x)$ заданной точности, является делом малоперспективным, особенно для методов более-менее высоких порядков. Поэтому главным способом отслеживания точности при реализации численных процессов решения задачи Коши остается применение различных полуэмпирических правил, основанных на принципе Рунге (см. § 12.5).

Будем считать, что при использовании метода p-го порядка абсолютная шаговая погрешность должна находиться в пределах $\varepsilon > 0$. Тогда, согласно ***принципу Рунге***, осуществляется счет по

системе узлов $x_i(h)=x_0+ih$ с шагом h и по системе узлов $x_j\left(\frac{h}{2}\right)=x_0+j\frac{h}{2}$ с шагом $\frac{h}{2}$. При четных j вторая система будет совпадать с первой, т.е. $x_i(h)=x_{2i}\left(\frac{h}{2}\right)$. Переход от расчетной точки x_i с приближенным значением решения в ней y_i к расчетной точке x_{i+1} один раз совершается за один шаг длины h и приводит к значению $y_{i+1}(h)\approx y(x_{i+1}(h))$, другой раз — за два шага длины $\frac{h}{2}$ («транзитом» через точку $x_i+\frac{h}{2}=x_{2i+1}\left(\frac{h}{2}\right)$ со значением $y_{2i+1}\left(\frac{h}{2}\right)\approx y\left(x_i+\frac{h}{2}\right)$) и дает значение

$$y_{2i+2}\left(\frac{h}{2}\right)\approx y\left(x_{2i+2}\left(\frac{h}{2}\right)\right)=y(x_{i+1}(h)).$$

Поправка Ричардсона в таком случае будет составлять величину

$$R_i\left(\frac{h}{2}\right):=\frac{y_{2i+2}\left(\frac{h}{2}\right)-y_{i+1}(h)}{2^p-1}. \tag{14.34}$$

Если величина $\left|R_i\left(\frac{h}{2}\right)\right|$ меньше заданного ε, то можно считать, что ошибка приближенного равенства $y(x_{i+1})\approx y_{2i+2}\left(\frac{h}{2}\right)$ не превосходит ε. Если же $\left|R_i\left(\frac{h}{2}\right)\right|>\varepsilon$, то следует уменьшить расчетный шаг h. При условии $\left|R_i\left(\frac{h}{2}\right)\right|<<\varepsilon$ стоит попытаться двигаться дальше с более крупным шагом (например, удвоить h).

Пример 14.2 (продолжение примера 14.1, см. § 14.3).

Посмотрим, что дает применение принципа Рунге к нескольким простым методам численного решения того же уравнения $y'=2x-3y$ с начальным условием $y(0)=1$. Из точки $x=0$ перейдем в точку $x=0.2$ за один шаг $h=0.2$ четырьмя одношаговыми методами: явным и неявным

методами Эйлера, методом трапеций и методом Хойна — частным случаем метода Рунге–Кутты второго порядка. С помощью полученных значений $y_1^{h=0.2}\approx y(0.2)$ и найденных ранее теми же методами в примере 14.1 значений $y_2^{h=0.1}\approx y(0.2)$ подсчитаем поправки Ричардсона

$$R_1^{h=0.1}=\frac{y_2^{h=0.1}-y_1^{h=0.2}}{2^p-1}$$

при $p=1$ для методов Эйлера и $p=2$ для методов трапеций и Хойна. Эти результаты, а также уточненные прибавлением к значениям $y_2^{h=0.1}$ поправок Ричардсона $R_1^{h=0.1}$ приближенные значения $\tilde{y}_2^{h=0.1}$ решения $y(0.2)$ и их истинные погрешности $y(0.2)-\tilde{y}_2^{h=0.1}$ сведем в следующую таблицу.

Метод	$y_1^{h=0.2}$	$R_1^{h=0.1}$	$\tilde{y}_2^{h=0.1}$	$y(0.2)-\tilde{y}_2^{h=0.1}$	$y(0.2)-y_2^{h=0.1}$
Эйлера (14.8)	0.4	0.11	0.62	≈ -0.04	≈ 0.07
Неявный Эйлера (14.14)	0.675	≈ -0.0407	≈ 0.5936	≈ -0.01	≈ -0.05
Трапеций (14.15)	≈ 0.5692	≈ 0.0032	≈ 0.5820	≈ -0.0001	≈ 0.003
Хойна (14.16)	0.62	≈ -0.0102	≈ 0.5793	≈ 0.0026	≈ -0.008

В эту таблицу последним столбцом помещен последний столбец из таблицы результатов примера 14.1, содержащий погрешности значений $y_2^{h=0.1}$. Сравнение с ним столбца со значениями поправок Ричардсона показывает, что эти поправки хорошо отражают поведение погрешностей методов (хотя и не дают основания считать их модули оценками погрешностей), а предпоследнего — эффективность уточнения по правилу Рунге–Ричардсона.

Грубо обозначенная здесь технология пошагового контроля точности численного интегрирования дифференциальных уравнений и автоматического выбора расчетного шага при этом на основе двойного счета в такой непосредственной форме говорит о ее значительной «дороговизне». Действительно, предположим, что для решения задачи (14.1)–(14.2) применяется четырехэтапный метод Рунге–Кутты четвертого порядка (14.33). Тогда вы-

полнение одного его шага с контролем точности по правилу Рунге потребует 11 вычислений правой части уравнения (14.1) (по четыре для получения каждого из значений $y_{i+1}(h)$, $y_{2i+1}\left(\frac{h}{2}\right)$ и $y_{2i+2}\left(\frac{h}{2}\right)$ минус одно общее для $y_{i+1}(h)$ и $y_{2i+1}\left(\frac{h}{2}\right)$), что весьма затратно.

Более «дешевый», но, возможно, менее строгий способ судить о том, достаточно ли малым выбран шаг h расчетов по методу Рунге–Кутты четвертого порядка (14.33), — это вычисление при каждом $i = 0, 1, 2, \ldots$ величин

$$\Theta_i = \left|\frac{\eta_2^i - \eta_3^i}{\eta_1^i - \eta_2^i}\right|.$$

Считается, что если величина Θ_i не превосходит нескольких сотых, то можно продолжить вычисления с данным шагом или пытаться при переходе от i к $i+1$ его увеличить; в противном случае шаг следует уменьшить, например, вдвое [62, 89, 188 и др.].

Стремление повысить вычислительную эффективность привело к появлению различных вычислительных версий методов Рунге–Кутты, благо для этого в семействе методов (14.32) имеется значительное число свободных параметров. Основные соображения, положенные в основу этих версий, таковы: нужно получить формулы из семейства методов Рунге–Кутты (14.32), которые использовали бы одни и те же значения функции — правой части уравнения (14.1) — и определяли бы разные конкретные методы одного порядка (или смежных порядков, например, четвертого и пятого); при этом, чтобы по разности результатов подсчета приближенных значений решения по выведенным близким формулам (с одним и тем же шагом h(!)) можно было судить о точности одного из них.

Приведем один из таких методов, который называется ***методом Кутты–Мерсона*** или, иначе, ***пятиэтапным методом Рунге–Кутты четвертого порядка***, а также ***методом вложенных форм*** [198].

На i-ом шаге решения задачи (14.1)–(14.2) последовательно вычисляют:

$$\begin{aligned}
\eta_1^i &= f(x_i, y_i),\\
\eta_2^i &= f\left(x_i + \frac{h}{3}, y_i + \frac{h}{3}\eta_1^i\right),\\
\eta_3^i &= f\left(x_i + \frac{h}{3}, y_i + \frac{h}{6}\eta_1^i + \frac{h}{6}\eta_2^i\right),\\
\eta_4^i &= f\left(x_i + \frac{h}{2}, y_i + \frac{h}{8}\eta_1^i + \frac{3h}{8}\eta_2^i\right),\\
\eta_5^i &= f\left(x_i + h, y_i + \frac{h}{2}\eta_1^i - \frac{3h}{2}\eta_3^i + 2h\eta_4^i\right),\\
\widetilde{y}_{i+1} &= y_i + \frac{h}{2}\left(\eta_1^i - 3\eta_3^i + 4\eta_4^i\right),\\
y_{i+1} &= y_i + \frac{h}{6}\left(\eta_1^i + 4\eta_4^i + \eta_5^i\right).
\end{aligned}$$

После этого подсчитывают величину

$$R := 0.2\left|y_{i+1} - \widetilde{y}_{i+1}\right|$$

и проводят сравнения. Если значение R окажется больше заданного допустимого уровня абсолютных погрешностей ε, то шаг уменьшают вдвое $\left(h := \frac{h}{2}\right)$ и возвращаются к началу второго этапа, т.е. заново вычисляют η_2^i, η_3^i и т.д. Если $R \le \varepsilon$, то считают $y(x_{i+1}) :\approx y_{i+1}$ с точностью ε. При переходе к следующему шагу делается проверка на возможность увеличить расчетный шаг: если $R \le \frac{\varepsilon}{64}$, то далее расчет ведется с шагом $h := 2h$.

Другие методы подобного типа можно найти, например, в монографии [6], где кроме приведенной модификации Мерсона метода Рунге–Кутты содержатся описание и расчетные формулы ***модификаций Фельберга*** и ***Инглენда***.

Замечание 14.1. Осуществить ***апостериорный контроль глобальной погрешности***, т.е. погрешности последнего вычисленного тем или иным шаговым методом значения $y_n \approx y(x_n) = y(b)$, можно, например, следующим образом [62].

Из интегрального равенства (14.3) при $x = x_n$ имеем

$$y(x_n) = y_0 + \int_{x_0}^{x_n} f(x, y(x))dx. \qquad (14.35)$$

Считая, что численным методом получена достаточно густая равномерная сеть приближенных значений y_i решения $y(x)$ задачи (14.1)–(14.2) и, следовательно, известны значения $f_i := f(x_i, y_i) \approx f(x_i, y(x_i))$, можно применить какую-либо квадратурную формулу замкнутого типа для вычисления фигурирующего в равенстве (14.35) интеграла (в частности, для этих целей можно использовать алгоритм Ромберга, см. гл.5). Величина разности между y_n и значением выражения $y_0 + \sum_i A_i f_i$, где A_i — коэффициенты используемой квадратурной формулы, позволяет приближенно судить о точности значения y_n. Оснований для этого тем больше, чем выше точность выполняемой для такого приближенного глобального контроля квадратуры.

УПРАЖНЕНИЯ

14.1. Последовательным дифференцированием уравнения

$$y' = xy + 1$$

найдите его приближенное частное решение, соответствующее начальному условию $y(0) = 0$, в виде отрезка степенного ряда до пятой степени включительно.

Можно ли получить тот же отрезок ряда методом последовательных приближений Пикара? Реализуйте пять итераций по формуле (14.5).

14.2. Запишите совокупность формул, определяющих метод пошаговых последовательных приближений, основанный на применении к обобщающему (14.3) интегральному уравнению

$$y(x) = y(x_i) + \int_{x_i}^{x} f(t, y(t))dt$$

при каждом $x_i = x_0 + ih$ фиксированного числа k итераций по методу Пикара (за $y(x_i)$ в этом уравнении должно быть принято значение в точке x_i результата последней итерации).

Опробуйте такой метод на задаче предыдущего упражнения, варьируя h и k и добиваясь одинаковой точности результатов в точке $x = 2$, задавая эту точность самостоятельно.

14.3. Проверьте справедливость численных результатов, приведенных в таблицах примеров 14.1 и 14.2 (см. § 14.3 и § 14.14).

14.4. Убедитесь, что для линейного дифференциального уравнения с постоянными коэффициентами

$$y' = py + q \qquad (14.36)$$

результат применения метода Рунге–Кутты второго порядка (7.30) не зависит от выбора значения параметра α ($\neq 0$).

Справедливо ли подобное утверждение для линейных ОДУ с переменными коэффициентами?

14.5. Можно ли утверждать, что метод Рунге–Кутты второго порядка (14.30), примененный к уравнению (14.36), приведет к тем же результатам, что и исправленный метод Эйлера (14.25)?

14.6. На задаче

$$y' = y^2 e^x - 2y, \quad y(0) = 0.5$$

протестируйте методы Рунге–Кутты второго порядка (14.30) при $\alpha = \frac{1}{2}$, $\alpha = 1$ и каком-нибудь другом $\alpha \neq 0$, а также исправленный метод Эйлера (14.25), заполнив таблицу по образцу сводной таблицы результатов примеров 14.1, 14.2. Как ведут себя здесь поправки Ричардсона (14.34) сравнительно с истинными погрешностями? Насколько уточняются приближенные решения в результате прибавления поправок Ричардсона?

14.7. Разработайте какой-либо конкретный алгоритм для реального воплощения описанного в замечании 14.1 (см. § 14.7) способа приближенного апостериорного оценивания глобальной погрешности результатов численного интегрирования начальной задачи (14.1)–(14.2) методом Рунге–Кутты второго порядка.

Проверьте работу этого алгоритма в условиях упражнения 14.6 на отрезке $[0, 1]$.

14.8. Запишите вид функции $\varphi(x, y, h)$, при котором равенство (14.26) определяет классический метод Рунге–Кутты четвертого порядка.

ГЛАВА 15 ЛИНЕЙНЫЕ МНОГОШАГОВЫЕ МЕТОДЫ

На интегро-интерполяционной основе выводится несколько семейств методов численного решения задачи Коши для ОДУ, имеющих перед многоэтапными одношаговыми методами таких же порядков точности то преимущество, что требуют меньших вычислительных затрат, благодаря использованию информации о решении в нескольких предыдущих точках. Показывается, как совместное применение явных и неявных формул одного порядка точности можно использовать для простого приближенного пошагового учета погрешностей. Наряду с конкретными семействами многошаговых методов, таких как методы Адамса и Милна, рассматривается общая структура многошаговых методов; демонстрируется возможность фиксирования параметров таких методов, исходя из понятия алгебраического порядка точности метода. Отмечается применимость всех рассматриваемых методов к решению систем дифференциальных уравнений и уравнений высших порядков. Описано построение методов Адамса и Адамса–Штёрмера для уравнений второго порядка.

15.1. МНОГОШАГОВЫЕ МЕТОДЫ АДАМСА

Как и в предыдущей главе, будем строить численные методы решения начальной задачи

$$y' = f(x, y), \quad x \in [x_0, b], \tag{15.1}$$

$$y(x_0) = y_0. \tag{15.2}$$

Будем считать, что уже найдено несколько приближенных значений $y_j \approx y(x_j)$ $(j = 0, 1, \ldots, i)$ решения $y = y(x)$ задачи (15.1)–(15.2) на равномерной сетке $x_j = x_0 + jh$, и нужно получить правило для вычисления очередного значения $y_{i+1} \approx y(x_{i+1})$. Для вывода таких правил используем интегро-интерполяционный подход. А именно, проинтегрировав левую и правую части уравнения (15.1) по промежутку $[x_i, x_{i+1}]$, в полученном равенстве

$$y(x_{i+1}) = y(x_i) + \int_{x_i}^{x_{i+1}} f(x, y(x))dx \tag{15.3}$$

под интеграл вместо функции $f(x, y(x))$ подставим интерполирующий ее многочлен $P_k(x)$. Хотя выражение функции $f(x, y(x))$, как функции одной переменной x, вообще говоря, неизвестно, ее дискретные приближенные значения $f(x_j, y_j) \approx f(x_j, y(x_j))$, обозначаемые в дальнейшем для краткости f_j, при $j = 1, 2, \ldots, i$ можно считать известными. В таком случае, дополняя эти известные значения пока что неизвестным значением $f_{i+1} := f(x_{i+1}, y_{i+1})$, можно построить таблицу конечных разностей (табл.15.1), служащую основой для образования интерполяционных многочленов k-й степени для интерполирования назад из точек (x_i, f_i) и (x_{i+1}, f_{i+1}).

Таблица 15.1

Таблица конечных разностей для построения конечноразностных формул Адамса

x_j	f_j	Δf_j	$\Delta^2 f_j$	$\Delta^3 f_j$	…	$\Delta^k f_j$
x_{i-k}	f_{i-k}					
		Δf_{i-k}				
x_{i-k+1}	f_{i-k+1}		$\Delta^2 f_{i-k}$			
		Δf_{i-k+1}		$\Delta^3 f_{i-k}$		
x_{i-k+2}	f_{i-k+2}		$\Delta^2 f_{i-k+1}$			
		Δf_{i-k+2}				
x_{i-k+3}	f_{i-k+3}					$\Delta^k f_{i-k}$
…	…	…	…	…	…	$\Delta^k f_{i-k+1}$
x_{i-3}	f_{i-3}					
		Δf_{i-3}	$\Delta^2 f_{i-3}$			
x_{i-2}	f_{i-2}			$\Delta^3 f_{i-3}$		
		Δf_{i-2}	$\Delta^2 f_{i-2}$			
x_{i-1}	f_{i-1}			$\Delta^3 f_{i-2}$		
		Δf_{i-1}	$\Delta^2 f_{i-1}$			
x_i	f_i					
		Δf_i				
x_{i+1}	f_{i+1}					

При интерполировании назад из узла x_i по второй интерполяционной формуле Ньютона (8.28) имеем

$$P_k(x) = P_k(x_i + qh) = f_i + q\Delta f_{i-1} + \frac{q(q+1)}{2!}\Delta^2 f_{i-2} +$$

$$+ \frac{q(q+1)(q+2)}{3!}\Delta^3 f_{i-3} + \ldots + \frac{q(q+1)\ldots(q+k-1)}{k!}\Delta^k f_{i-k} \tag{15.4}$$

(см. конечные разности, подчеркнутые в табл.15.1 сплошной линией), а из узла x_{i+1} по той же формуле получаем многочлен

$$\widetilde{P}_k(x) = \widetilde{P}_k(x_{i+1}+qh) = f_{i+1} + q\Delta f_i + \frac{q(q+1)}{2!}\Delta^2 f_{i-1} +$$

$$+\frac{q(q+1)(q+2)}{3!}\Delta^3 f_{i-2} + \ldots + \frac{q(q+1)\ldots(q+k-1)}{k!}\Delta^k f_{i-k+1} \quad (15.5)$$

(использующий разности, подчеркнутые пунктиром).

Подстановка многочленов $P_k(x)$ и $\widetilde{P}_k(x)$ в равенство (15.3) приводит к формулам для вычисления очередного значения $y_{i+1} \approx y(x_{i+1})$ вида

$$y_{i+1} = y_i + \int_{x_i}^{x_{i+1}} P_k(x)dx \quad (15.6)$$

и

$$y_{i+1} = y_i + \int_{x_i}^{x_{i+1}} \widetilde{P}_k(x)dx. \quad (15.7)$$

В результате применения к интегралам в (15.6) и (15.7) формулы Ньютона–Лейбница получается два семейства методов (с параметром $k \in \mathbf{N}_0$), называемых ***многошаговыми методами Адамса****). Рассмотрим по отдельности каждое из этих семейств.

Экстраполяционные методы Адамса–Башфорта. Чтобы подставить в (15.6) многочлен (15.4), зависящий от переменной $q = \dfrac{x - x_i}{h}$, сделаем в интеграле $\int_{x_i}^{x_{i+1}} P_k(x)dx$ замену переменной $x = x_i + qh$, в соответствии с которой

$$\int_{x_i}^{x_{i+1}} P_k(x)\,dx = h\int_0^1 P_k(x_i+qh)\,dq.$$

Тогда формула (15.6) может быть переписана в виде

$$y_{i+1} = y_i + hI_k, \quad (15.8)$$

*) А́дамс Джон Кауч (1819–1892) — английский математик и астроном. Одновременно с Леверье им было предсказано существование планеты Нептун.

где

$$I_k := \int_0^1 P_k(x_i+qh)\,dq = \left[f_i q + \frac{q^2}{2}\Delta f_{i-1} + \left(\frac{q^3}{6}+\frac{q^2}{4}\right)\Delta^2 f_{i-2} + \right.$$

$$\left. +\frac{1}{6}\left(\frac{q^4}{4}+q^3+q^2\right)\Delta^3 f_{i-3} + \frac{1}{24}\left(\frac{q^5}{5}+\frac{3q^4}{2}+\frac{11q^3}{3}+3q^2\right)\Delta^4 f_{i-4}+\ldots\right]_0^1 =$$

$$= f_i + \frac{1}{2}\Delta f_{i-1} + \frac{5}{12}\Delta^2 f_{i-2} + \frac{3}{8}\Delta^3 f_{i-3} + \frac{251}{720}\Delta^4 f_{i-4} + \ldots. \quad (15.9)$$

Таким образом, на основе (15.8) получается следующая конечно-разностная формула, определяющая ***экстраполяционный метод Адамса–Башфорта***:

$$y_{i+1} = y_i + h\left(f_i + \frac{1}{2}\Delta f_{i-1} + \frac{5}{12}\Delta^2 f_{i-2} + \frac{3}{8}\Delta^3 f_{i-3} + \frac{251}{720}\Delta^4 f_{i-4} + \ldots\right). \quad (15.10)$$

Посмотрим, что представляют собой наиболее простые **частные случаи метода Адамса–Башфорта**, соответствующие нескольким первым значениям параметра k в формуле (15.8). Сразу заметим, что при фиксировании $k = 0, 1, 2, \ldots$ в (15.8) тем самым задается степень интерполяционного многочлена (нулевая, первая, вторая и т.д.) и, соответственно, число слагаемых, равное 1, 2, 3, …, в правой части (15.9) (или, что то же, в скобках формулы (15.10)). Конечные разности в получающихся при этом конкретных формулах будем раскрывать через значения функции, приводя формулы к виду, называемому иногда ***ординатным***. Имеем:

при $k = 0$

$$I_0 = f_i \quad \Rightarrow \quad y_{i+1} = y_i + hf(x_i, y_i); \quad (15.11)$$

при $k = 1$

$$I_1 = f_i + \frac{1}{2}\Delta f_{i-1} = \frac{3}{2}f_i - \frac{1}{2}f_{i-1} \quad \Rightarrow$$

$$y_{i+1} = y_i + \frac{h}{2}[3f(x_i, y_i) - f(x_{i-1}, y_{i-1})]; \quad (15.12)$$

при $k = 2$

$$I_2 = I_1 + \frac{5}{12}\Delta^2 f_{i-2} = \frac{23}{12} f_i - \frac{16}{12} f_{i-1} + \frac{5}{12} f_{i-2} \quad \Rightarrow$$

$$y_{i+1} = y_i + \frac{h}{12}[23f(x_i, y_i) - 16f(x_{i-1}, y_{i-1}) + 5f(x_{i-2}, y_{i-2})]; \quad (15.13)$$

при $k = 3$

$$I_3 = I_2 + \frac{3}{8}\Delta^3 f_{i-3} = \frac{55}{24} f_i - \frac{59}{24} f_{i-1} + \frac{37}{24} f_{i-2} - \frac{9}{24} f_{i-3} \quad \Rightarrow$$

$$y_{i+1} = y_i + \frac{h}{24}[55f(x_i, y_i) - 59f(x_{i-1}, y_{i-1}) + 37f(x_{i-2}, y_{i-2}) - 9f(x_{i-3}, y_{i-3})]. \quad (15.14)$$

Формулы (15.11), (15.12), (15.13) *и* (15.14) *определяют методы Адамса–Башфорта соответственно первого, второго, третьего и четвертого порядков.* Относительно порядка метода (15.11) сомнений нет: мы узнаём ***метод Эйлера*** (14.8), вопрос о порядке которого обсуждался в § 14.2. Обоснование же утверждения о порядках остальных перечисленных конкретных методов состоит в следующем.

В общем случае, для $k+1$ раз непрерывно дифференцируемой функции шаговая ошибка может быть получена сравнением равенства (15.6) с породившим его равенством (15.3), т.е. интегрированием остаточного члена $R_k(x) = \frac{f^{(k+1)}(\xi, y(\xi))}{(k+1)!}\Pi_{k+1}(x)$ интерполяционной формулы Лагранжа (см.(8.12)). Применительно к конечноразностной интерполяционной формуле (15.4) функция $R_k(x)$ преобразуется к виду

$$R_k(x_i + qh) = \frac{f^{(k+1)}(\xi, y(\xi))}{(k+1)!} q(q+1)\dots(q+k)h^{k+1}, \quad (15.15)$$

т.е. может считаться величиной $O(h^{k+1})$. Следовательно, локальная погрешность метода типа (15.8) будет составлять величину $h\int_0^1 R_k(x_i + qh)dq = O(h^{k+2})$, а глобальная — величину $O(h^{k+1})$.

Таким образом, *метод Адамса, порождаемый интерполированием правой части уравнения* (15.1) *с помощью многочлена k-й степени, является методом* $(k+1)$*-го порядка точности* (относительно шага h).

Так как интерполяционный многочлен $P_k(x)$ степени k строился по $k+1$ значениям $f_{i-k}, f_{i-k+1}, \dots, f_{i-1}, f_i$, использующим, в свою очередь, $k+1$ значений (вообще говоря, приближенных) $y_{i-k}, y_{i-k+1}, \dots, y_{i-1}, y_i$ искомого решения $y(x)$, а количество этих значений (даже если какие-то значения между y_k и y_i явно не участвуют) определяет ***шаговость метода,*** то для методов семейства (15.10) мы наблюдаем совпадение порядка метода с тем, сколькишаговым он является. Следовательно, метод (15.12) — это двухшаговый метод Адамса–Башфорта второго порядка, метод (15.13) — трехшаговый метод Адамса–Башфорта третьего порядка, и т.д.

Название ***экстраполяционный метод*** связано с тем, что интерполяционный многочлен $P_k(x)$ для равенства (15.6) строился по узлам, расположенным на промежутке $[x_{i-k}, x_i]$, а применялся к отрезку $[x_i, x_{i+1}]$, т.е. производилось интерполирование в широком смысле, а в сущности делалось экстраполирование (по поводу нюансов этой терминологии см. сноски в § 8.2). Понятно, что в названии рассматриваемого далее второго семейства методов Адамса слово «интерполяционный» употребляется в узком смысле.

Интерполяционные методы Адамса–Моултона. В интеграле, фигурирующем в формуле (15.7), делаем замену $x = x_{i+1} + qh$ и подставляем в него выражение $\widetilde{P}_k(x)$, определяемое формулой (15.5). Приходим к аналогичному (15.8) равенству

$$y_{i+1} = y_i + h\widetilde{I}_k,$$

где

$$\widetilde{I}_k := \int_{-1}^{0}\widetilde{P}_k(x_{i+1} + qh)\,dq = \left[f_{i+1}q + \frac{q^2}{2}\Delta f_i + \left(\frac{q^3}{6} + \frac{q^2}{4}\right)\Delta^2 f_{i-1} + \right.$$

$$\left. + \frac{1}{6}\left(\frac{q^4}{4} + q^3 + q^2\right)\Delta^3 f_{i-2} + \frac{1}{24}\left(\frac{q^5}{5} + \frac{3q^4}{2} + \frac{11q^3}{3} + 3q^2\right)\Delta^4 f_{i-3} + \dots \right]_{-1}^{0} =$$

$$= f_{i+1} - \frac{1}{2}\Delta f_i - \frac{1}{12}\Delta^2 f_{i-1} - \frac{1}{24}\Delta^3 f_{i-2} - \frac{19}{720}\Delta^4 f_{i-3} - \dots . \quad (15.16)$$

Отсюда следует конечноразностная формула ***интерполяционного метода Адамса–Моултона***

$$y_{i+1} = y_i + h\left[f_{i+1} - \frac{1}{2}\Delta f_i - \frac{1}{12}\Delta^2 f_{i-1} - \right.$$

$$\left. - \frac{1}{24}\Delta^3 f_{i-2} - \frac{19}{720}\Delta^4 f_{i-3} - \ldots \right]. \quad (15.17)$$

Аналогично тому, как это делалось для методов Адамса–Башфорта, при $k = 0, 1, 2, 3$, т.е. фиксированием одного, двух, трех, четырех членов в представлении (15.16) интеграла $\tilde{I}_k(x)$, получаем следующие частные формулы:

при $k = 0$

$$y_{i+1} = y_i + hf(x_{i+1}, y_{i+1}), \quad (15.18)$$

при $k = 1$

$$y_{i+1} = y_i + \frac{h}{2}[f(x_{i+1}, y_{i+1}) + f(x_i, y_i)], \quad (15.19)$$

при $k = 2$

$$y_{i+1} = y_i + \frac{h}{12}[5f(x_{i+1}, y_{i+1}) + 8f(x_i, y_i) - f(x_{i-1}, y_{i-1})], \quad (15.20)$$

при $k = 3$

$$y_{i+1} = y_i + \frac{h}{24}[9f(x_{i+1}, y_{i+1}) + 19f(x_i, y_i) -$$

$$- 5f(x_{i-1}, y_{i-1}) + f(x_{i-2}, y_{i-2})]. \quad (15.21)$$

Формулы (15.18) и (15.19) определяют уже известные нам методы, а именно, ***неявный метод Эйлера*** (14.14) и ***метод трапеций*** (14.15), имеющие первый и второй порядки точности соответственно. Заметим, что оба эти метода являются одношаговыми, а следующие за ними методы Адамса–Моултона (15.20) и (15.21) третьего и четвертого порядков относятся, как легко видеть, соответственно к двухшаговым и трехшаговым методам. Таким образом, *для интерполяционных методов Адамса–Моултона порядок шаговости на единицу ниже порядка точности метода* (за тривиальным исключением, отвечающим случаю $k = 0$).

Важное различие в экстраполяционных и интерполяционных методах Адамса заключается в том, что первые из них являются явными, а вторые — неявными. Эти термины однозначно определяют, о каком из двух семейств методов Адамса идет речь, а их сущность диктует особенности использования методов Адамса при практических расчетах, что найдет отражение в следующем параграфе.

15.2. МЕТОДЫ ПРОГНОЗА И КОРРЕКЦИИ. ПРЕДИКТОР-КОРРЕКТОРНЫЕ МЕТОДЫ АДАМСА

Под названием ***методы прогноза и коррекции*** (иначе ***методы предсказания и уточнения, предиктор-корректорные методы***) понимается совместное применение явных и неявных методов одинакового или смежных порядков. По явной формуле значение решения $y(x)$ задачи (15.1)–(15.2) в текущей (расчетной) точке x_{i+1} прогнозируется, т.е. находится его, быть может, достаточно грубое приближение, а с помощью неявной формулы, в правую часть которой подставляется спрогнозированное значение, оно уточняется (корректируется). Пример приближенного вычисления $y(x_{i+1})$ по такой явно-неявной схеме у нас уже есть: в § 14.3 рассматривалось парное использование явного метода Эйлера для предсказания и метода трапеций для уточнения (см. итерационную формулу (14.17), определяющую усовершенствованный метод Эйлера–Коши с итерационной обработкой).

Остановимся подробнее на методах прогноза и коррекции, базирующихся на парах явных и неявных методов Адамса одинакового порядка. Обозначим через $y_{i+1}^{Б}$ приближенное значение решения $y(x_{i+1})$, подсчитываемое по явной экстраполяционной формуле Адамса–Башфорта, и составим несколько пар из рассмотренных в предыдущем параграфе частных формул Адамса–Башфорта (15.11), (15.12), (15.13), (15.14) и Адамса–Моултона (15.18), (15.19), (15.20), (15.21).

Имеем следующие ***предиктор-корректорные методы Адамса***:

первого порядка (он же ***явно-неявный метод Эйлера***)

$$\begin{cases} y_{i+1}^{Б} = y_i + hf(x_i, y_i), \\ y_{i+1} = y_i + hf\left(x_{i+1}, y_{i+1}^{Б}\right); \end{cases}$$

второго порядка

$$\begin{cases} y_{i+1}^{Б} = y_i + \frac{h}{2}[3f(x_i, y_i) - f(x_{i-1}, y_{i-1})], \\ y_{i+1} = y_i + \frac{h}{2}\left[f\left(x_{i+1}, y_{i+1}^{Б}\right) + f(x_i, y_i)\right]; \end{cases}$$

третьего порядка

$$\begin{cases} y_{i+1}^{Б} = y_i + \frac{h}{12}[23f(x_i, y_i) - 16f(x_{i-1}, y_{i-1}) + 5f(x_{i-2}, y_{i-2})], \\ y_{i+1} = y_i + \frac{h}{12}\left[5f\left(x_{i+1}, y_{i+1}^{Б}\right) + 8f(x_i, y_i) - f(x_{i-1}, y_{i-1})\right]; \end{cases}$$

четвертого порядка

$$\begin{cases} y_{i+1}^{Б} = y_i + \frac{h}{24}[55f(x_i, y_i) - 59f(x_{i-1}, y_{i-1}) + \\ \qquad\qquad + 37f(x_{i-2}, y_{i-2}) - 9f(x_{i-3}, y_{i-3})], \\ y_{i+1} = y_i + \frac{h}{24}[9f\left(x_{i+1}, y_{i+1}^{Б}\right) + 19f(x_i, y_i) - \\ \qquad\qquad - 5f(x_{i-1}, y_{i-1}) + f(x_{i-2}, y_{i-2})]. \end{cases} \quad (15.22)$$

Одним из главных достоинств методов прогноза и коррекции является возможность контролировать шаговую погрешность сравнением двух полученных по явной и неявной формулам приближений к $y(x_{i+1})$. Покажем, как реализуется эта возможность для наиболее употребительного предиктор-корректорного метода Адамса четвертого порядка (15.22).

Вспомним, что первая из формул (15.22) была получена из общей формулы Адамса–Башфорта (15.10), а вторая — из общей формулы Адамса–Моултона (15.17), в которых последними брались разности третьего порядка (подынтегральная функция в равенстве (15.3) аппроксимировалась интерполяционным многочленом третьей степени). Считая, что расчетный шаг h достаточно мал и конечные разности с ростом их порядка убывают, главные части шаговых погрешностей формул Башфорта и Моултона четвертого порядка, в соответствии с (15.10) и (15.17), характеризуются величинами $\frac{251}{720}h\Delta^4 f_{i-4}$ для явной и $-\frac{19}{720}h\Delta^4 f_{i-3}$ для неявной формул. Таким образом, если наряду с введенным обозначением $y_{i+1}^{Б}$ обозначить через y_{i+1}^{M} приближенное значение $y(x_{i+1})$, получаемое по формуле Адамса–Моултона четвертого порядка, то можно записать два приближенных представления $y(x_{i+1})$:

$$y(x_{i+1}) \approx y_{i+1}^{Б} + \frac{251}{720}h\Delta^4 f_{i-4} \quad (15.23)$$

и

$$y(x_{i+1}) \approx y_{i+1}^{M} - \frac{19}{720}h\Delta^4 f_{i-3}. \quad (15.24)$$

Отсюда видно, что если четвертые разности функции $f(x, y(x))$ в используемой части табл.15.1 конечных разностей практически постоянны (а это можно связать с удачным выбором величины шага h при достаточном запасе знаков в значениях $f(x_j, y_j)$), то, во-первых, значения $y_{i+1}^{Б}$ и y_{i+1}^{M} дают двусторонние приближения к точному решению $y(x_{i+1})$, а во-вторых, через разность между значениями $y_{i+1}^{Б}$ и y_{i+1}^{M} можно оценить точность каждого из них.

Действительно, приравнивая правые части приближенных равенств (15.23) и (15.24) и отождествляя $\Delta^4 f_{i-4}$ с $\Delta^4 f_{i-3}$, имеем:

$$y_{i+1}^{M} - y_{i+1}^{Б} \approx \frac{19}{720}h\Delta^4 f_{i-3} + \frac{251}{720}h\Delta^4 f_{i-4} \approx \frac{3}{8}h\Delta^4 f_{i-3},$$

откуда

$$h\Delta^4 f_{i-3} \approx \frac{8}{3}\left(y_{i+1}^{M} - y_{i+1}^{Б}\right).$$

Подставляя последнее в (15.24), получаем приближенное равенство

$$y(x_{i+1}) \approx y_{i+1}^{M} - \frac{19}{270}\left(y_{i+1}^{M} - y_{i+1}^{Б}\right). \quad (15.25)$$

Использование приближенной формулы (15.25) может быть двояким. Переписав ее в виде

$$y(x_{i+1}) - y_{i+1}^{M} \approx -\frac{19}{270}\left(y_{i+1}^{M} - y_{i+1}^{Б}\right),$$

применяем это для пошагового контроля точности:

если $\frac{19}{270}\left|y_{i+1}^{M} - y_{i+1}^{Б}\right| < \varepsilon$, *то полагаем* $y(x_{i+1}) :\approx y_{i+1}^{M}$ *с точностью* ε ***и переходим*** *к следующему шагу* ($i := i+1$), *иначе уменьшаем шаг* h *и снова подсчитываем* y_{i+1}^{M} *и* $y_{i+1}^{Б}$.

Другое назначение формулы (15.25) — это прямое применение ее правой части для получения уточненного значения:

полагаем

$$y_{i+1} :\approx y_{i+1}^{M} - \frac{19}{270}\left(y_{i+1}^{M} - y_{i+1}^{Б}\right). \quad (15.26)$$

Наверное, есть смысл контроль точности делать на каждом шаге, а к уточнению по формуле (15.26) прибегать при выводе окончательных результатов.

Замечание 15.1. При выводе формулы (15.25) под y_{i+1}^{M} мы понимаем значение, соответствующее «чистому» методу Адамса–Моултона четвертого порядка, т.е. y_{i+1}^{M} — это точная реализация неявной формулы (15.21). Вторая же формула предиктор-корректорного метода (15.22) соответствует лишь одному приближению к y_{i+1}^{M} по методу простых итераций, где в качестве начального приближения берется $y_{i+1}^{Б}$. Поэтому примене-

ния формулы (15.25) к методу прогноза и коррекции (15.22) будут убедительны в том случае, если его вторая формула итерируется хотя бы один-два раза. Однако, чем больше таких итераций, тем ниже вычислительная эффективность этого метода, в целом весьма высокая по сравнению с многоэтапными методами Рунге–Кутты.

15.3. МЕТОД МИЛНА ЧЕТВЕРТОГО ПОРЯДКА

Рассмотрим еще один широко известный метод прогноза и коррекции — ***метод Милна***.

Для вывода первой формулы Милна (т.е. формулы предсказания) проинтегрируем уравнение (15.1) на промежутке $[x_{i-3}, x_{i+1}]$ и в полученном интегральном равенстве

$$y(x_{i+1}) = y(x_{i-3}) + \int_{x_{i-3}}^{x_{i+1}} f(x, y(x))\,dx \qquad (15.27)$$

подынтегральную функцию $f(x, y(x))$ заменим первым интерполяционным многочленом Ньютона $P_3(x)$, построенным по четырем узлам x_{i-3}, x_{i-2}, x_{i-1}, x_i с предполагающимися уже известными приближенными значениями

$$f_{i-3} := f(x_{i-3}, y_{i-3}) \approx f(x_{i-3}, y(x_{i-3})),$$
$$f_{i-2} := f(x_{i-2}, y_{i-2}) \approx f(x_{i-2}, y(x_{i-2})),$$
$$f_{i-1} := f(x_{i-1}, y_{i-1}) \approx f(x_{i-1}, y(x_{i-1})),$$
$$f_i := f(x_i, y_i) \approx f(x_i, y(x_i)).$$

Тогда, после замены переменной $x \approx x_{i-3} + qh$, на основании (15.27) имеем:

$$y(x_{i+1}) \approx y_{i-3} + h\int_0^4 P_3(x_{i-3}+qh)dq = y_{i-3} + h\int_0^4 \Big[f_{i-3} + q\Delta f_{i-3} +$$
$$+ \frac{q(q-1)}{2!}\Delta^2 f_{i-3} + \frac{q(q-1)(q-2)}{3!}\Delta^3 f_{i-3}\Big]dq = y_{i-3} +$$
$$+ h\left[f_{i-3}q + \frac{q^2}{2}\Delta f_{i-3} + \frac{1}{2}\left(\frac{q^3}{3} - \frac{q^2}{2}\right)\Delta^2 f_{i-3} + \frac{1}{6}\left(\frac{q^4}{4} - q^3 + q^2\right)\Delta^3 f_{i-3}\right]_0^4 =$$
$$= y_{i-3} + \frac{4}{3}h\left(3f_{i-3} + 6\Delta f_{i-3} + 5\Delta^2 f_{i-3} + 2\Delta^3 f_{i-3}\right).$$

Отсюда, выразив конечные разности через значения функции, получаем ***первую формулу Милна*** (предсказания)

$$\hat{y}_{i+1} = y_{i-3} + \frac{4}{3}h\left[2f(x_i, y_i) - f(x_{i-1}, y_{i-1}) + 2f(x_{i-2}, y_{i-2})\right], \qquad (15.28)$$

которую, очевидно, следует отнести к экстраполяционным.

Главный член локальной погрешности формулы (15.28) находим интегрированием следующего (первого из неучтенных) слагаемого интерполяционного многочлена Ньютона. Именно:

$$y(x_{i+1}) - \hat{y}_{i+1} \approx h\int_0^4 \frac{q(q-1)(q-2)(q-3)}{4!}\Delta^4 f_{i-3}dq =$$
$$= \frac{h}{24}\Delta^4 f_{i-3}\int_0^4 \left(q^4 - 6q^3 + 11q^2 - 6q\right)dq = \frac{14}{45}h\,\Delta^4 f_{i-3}.$$

Считая четвертые разности примерно одинаковыми, опустим индекс у функции f в записи $\Delta^4 f_{i-3}$; в результате получаем следующее приближенное представление решения в точке x_{i+1}:

$$y(x_{i+1}) \approx \hat{y}_{i+1} + \frac{14}{45}h\,\Delta^4 f. \qquad (15.29)$$

Вывод второй формулы Милна более прост. Проинтегрируем уравнение (15.1) теперь на промежутке $[x_{i-1}, x_{i+1}]$ и в полученном равенстве

$$y(x_{i+1}) = y(x_{i-1}) + \int_{x_{i-1}}^{x_{i+1}} f(x, y(x))dx$$

применим к интегралу простейшую формулу Симпсона (5.30). Имеем

$$y(x_{i+1}) = y(x_{i-1}) + \frac{h}{3}\left[f(x_{i+1}, y(x_{i+1})) + 4f(x_i, y(x_i)) +\right.$$
$$\left. + f(x_{i-1}, y(x_{i-1}))\right] - \frac{h^5}{90}f^{IV}(\xi_i). \quad (15.30)$$

Отбрасывая здесь остаточный член и заменяя значения решения $y(x_{i-1})$ и $y(x_i)$ известными приближенными значениями y_{i-1} и y_i, а стоящее в правой части под знаком функции f неизвестное значение $y(x_{i+1})$ тем значением $\hat{y}_{i+1}$, которое получается в результате вычислений по явной первой формуле Милна (15.28), приходим ко ***второй формуле Милна*** (уточнения)

$$y_{i+1} = y_{i-1} + \frac{h}{3}\left[f(x_{i+1}, \hat{y}_{i+1}) + 4f(x_i, y_i) + f(x_{i-1}, y_{i-1})\right], \quad (15.31)$$

являющейся интерполяционной.

Для вывода приближенной оценки шаговой погрешности воспользуемся приближенным равенством $f^{IV}(\xi) \approx \frac{\Delta^4 f}{h^4}$, где $\Delta^4 f$ так же, как и в (15.29), — условная запись практически постоянных четвертых разностей. Исходя из точного равенства (15.30), локальную погрешность получаемого с помощью формулы (15.31) (возможно, с итерационной обработкой, см. замечание 15.1) приближенного значения y_{i+1} можно приближенно охарактеризовать величиной $-\frac{h}{90}\Delta^4 f$, т.е.

$$y(x_{i+1}) \approx y_{i+1} - \frac{h}{90}\Delta^4 f. \qquad (15.32)$$

Сравнение (15.29) и (15.32) дает:

$$y_{i+1} - \hat{y}_{i+1} \approx 29\frac{h}{90}\Delta^4 f \quad \Rightarrow \quad \frac{h}{90}\Delta^4 f \approx \frac{y_{i+1} - \hat{y}_{i+1}}{29},$$

следовательно,

$$y(x_{i+1}) - y_{i+1} \approx \frac{\hat{y}_{i+1} - y_{i+1}}{29}. \qquad (15.33)$$

Таким образом, при численном интегрировании начальной задачи (15.1)–(15.2) методом Милна четвертого порядка, определенным формулами (15.28) и (15.31), на каждом i-м шаге следует вычислять величину

$$d_{i+1} := \frac{\hat{y}_{i+1} - y_{i+1}}{29}$$

и сравнивать ее модуль с величиной $\varepsilon > 0$ допустимой шаговой погрешности. Если $|d_{i+1}| < \varepsilon$, то за $y(x_{i+1})$ принимается полученное по второй формуле Милна значение y_{i+1} (или его уточненное значение $y_{i+1} := y_{i+1} + d_{i+1}$); иначе шаг должен быть уменьшен.

Фигурирующая в приближенном равенстве (15.33) постоянная $\frac{1}{29}$ примерно вдвое меньше постоянной $\frac{19}{270} \approx \frac{1}{14}$ в аналогичном равенстве (15.25) для предиктор-корректорного метода Адамса четвертого порядка (15.22), что характеризует метод Милна как несколько более точный при одинаковых вычислительных затратах. Но в дальнейшем (см. гл. 16, в частности, § 16.5) появятся и другие критерии для сравнения этих конкурирующих многошаговых методов.

15.4. ОБЩИЙ ВИД ЛИНЕЙНЫХ МНОГОШАГОВЫХ МЕТОДОВ. УСЛОВИЯ СОГЛАСОВАННОСТИ

Все m-шаговые методы Адамса можно описать одной формулой

$$y_{i+1} = y_i + h\sum_{j=0}^{m}\beta_j f(x_{i+1-j}, y_{i+1-j}), \qquad (15.34)$$

где m должно быть фиксированным натуральным числом, а i может принимать значения $m-1, m, m+1, \ldots$. При $\beta_0 = 0$ формула (15.34) определяет явные, а при $\beta_0 \neq 0$ — неявные методы, которые с единственными конкретными наборами коэффициентов $\beta_1, \beta_2, \beta_3, \ldots, \beta_m$ в первом и $\beta_0, \beta_1, \beta_2, \ldots, \beta_m$ во втором случаях являются соответственно экстраполяционными и интерполяционными методами Адамса. Как известно, m-шаговый метод Адамса при достаточной гладкости решения имеет m-й порядок точности. Так как при любом его порядке для реализации одного шага требуется вычисление лишь одного нового значения функции (или двух при парном применении явных и неявных методов Адамса, см. предыдущий параграф), то при построении решения на большом промежутке выгодно применять методы Адамса достаточно высоких порядков. Но при этом возникают проблемы с вычислением первых $m-1$ «разгонных» значений $y_1, y_2, y_3, \ldots, y_{m-1}$ (учитываем, что значение y_0 задано). Для их получения разработаны специальные процедуры разгона [6, 62, 84]. Впрочем, можно обойтись применением для этих целей одношаговых методов Рунге–Кутты.

Аналогично (15.34) при $m = 4$ выглядят и формулы Милна, только в формуле прогноза (15.28) первым слагаемым стоит y_{i-3}, а в формуле коррекции (15.31) — y_{i-1}. Похожий на (15.34) вид имеют также многошаговые методы Нистрёма, Хемминга и др. (см. [6]). Все эти методы можно описать формулой

$$y_{i+1} = \sum_{j=1}^{m}\alpha_j y_{i+1-j} + h\sum_{j=0}^{m}\beta_j f(x_{i+1-j}, y_{i+1-j}), \qquad (15.35)$$

которая задает ***общий вид линейных многошаговых методов*** (приближенное значение решения $y(x)$ в точке x_{i+1} представляется в виде линейной комбинации нескольких приближенных значений решения и его производной в этой и в предшествующих m точках).

К построению конкретных многошаговых методов, т.е. к фиксированию параметров α_j, β_j в формуле (15.35) при фиксированных значениях $m = 1, 2, 3, \ldots$ можно подойти следующим, отличным от рассмотренных ранее способом.

Обратившись к любой из выведенных в двух предыдущих

параграфах многошаговых формул вычисления приближенного значения y_{i+1} решения $y(x_{i+1})$ уравнения $y' = f(x, y)$, мы видим, что локальная ошибка, точнее, ее главная часть, представляется в виде (см. (15.23), (15.24), (15.29), (15.32))

$$y(x_{i+1}) - y_{i+1} = Ch\Delta^k f = C_1 h^{k+1} f^{(k)}(\xi_i) = C_1 h^{k+1} y^{(k+1)}(\xi_i), \quad (15.36)$$

где C и C_1 — некоторые постоянные, ξ_i — некоторая точка, а k — порядок метода. Так как многочлен k-й степени имеет нулевую $(k+1)$-ю производную, то, значит, метод получения y_{i+1}, для которого имеет место (15.36), точен, если его решение $y(x)$ — многочлен k-й (и менее) степени.

Подмеченный факт приводит к мысли положить в основу построения методов численного интегрирования задачи Коши (15.1)–(15.2) ***алгебраический порядок точности***, под которым здесь понимается максимальная степень k многочлена $y(x) = P_k(x)$, обращающего в нуль погрешность (15.36).

Будем подставлять в формулу (15.35), определяющую общий вид m-шаговых методов, в роли решения $y(x)$ представителей многочленов $P_k(x)$ — степенные функции $1, x, x^2, \ldots, x^k$, и будем считать, что такая подстановка вместо y_j значений $y(x_j)$ не порождает погрешностей, т.е. получающиеся при этом равенства являются точными.

Пусть $k = 0$, т.е. возьмем $y(x) \equiv 1$. Тогда, так как $y(x_j) = 1$ при любом значении j, а $y' = f(x, y) \equiv 0$, то (15.35) в этом случае превращается в равенство

$$1 = \sum_{j=1}^{m} \alpha_j \cdot 1 - h \sum_{j=0}^{m} \beta_j \cdot 0 .$$

Следовательно, параметры α_j любого из методов семейства (15.35) должны подчиняться условию

$$\sum_{j=1}^{m} \alpha_j = 1. \quad (15.37)$$

Положим теперь $k = 1$ и будем считать решением $y(x) \equiv x$. В силу $y' = f(x, y) \equiv 1$, результат подстановки такой функции в (15.35) приведет к равенству

$$x_{i+1} = \sum_{j=1}^{m} \alpha_j x_{i+1-j} + h \sum_{j=0}^{m} \beta_j , \quad (15.38)$$

которое должно быть справедливым при любых $i = m-1, m, m+1, \ldots$. Взяв $i = m$, исходя из выражения i-го узла

через шаг $x_i = x_0 + ih$, от (15.38) приходим к равенству

$$x_0 + (m+1)h = \sum_{j=1}^{m} \alpha_j [x_0 + (m+1-j)h] + h \sum_{j=0}^{m} \beta_j .$$

Переписав его в виде

$$x_0 + (m+1)h = x_0 \sum_{j=1}^{m} \alpha_j + (m+1)h \sum_{j=1}^{m} \alpha_j - h \sum_{j=1}^{m} j\alpha_j + h \sum_{j=0}^{m} \beta_j ,$$

учитываем найденное выше условие (15.37), в результате чего получаем простое требование ко второй группе параметров семейства m-шаговых методов (15.35):

$$\sum_{j=0}^{m} \beta_j = \sum_{j=1}^{m} j\alpha_j . \quad (15.39)$$

Совокупность требований (15.37), (15.39) называют ***условиями согласованности параметров*** линейных многошаговых методов (15.35). Нами они выведены как необходимые условия того, чтобы произвольный метод этого семейства имел, по меньшей мере, первый алгебраический порядок точности. Доказано [185], что они являются и достаточными для этого условиями.

Посмотрим, что представляют собой **одношаговые методы** из семейства m-шаговых методов, т.е. зафиксируем в (15.35) $m = 1$. Имеем трехпараметрическую формулу

$$y_{i+1} = \alpha_1 y_i + h(\beta_0 f_{i+1} + \beta_1 f_i). \quad (15.40)$$

Чтобы она определяла метод не ниже первого порядка, согласно условиям (15.37) и (15.39), должно быть

$$\alpha_1 = 1, \quad \beta_0 + \beta_1 = 1.$$

Таким образом, остается лишь одна степень свободы, т.е. на самом деле, (15.40) — это однопараметрическое семейство одношаговых методов

$$y_{i+1} = y_i + h[\beta_0 f_{i+1} + (1 - \beta_0) f_i]. \quad (15.41)$$

При любом значении β_0 метод (15.41) точен для многочленов первой степени. Например, при $\beta_0 = 0$ — это ***явный метод Эйлера***, а при $\beta_0 = 1$ — ***неявный метод Эйлера***. Оба они точны, если решением задачи (15.1)–(15.2) является линейная функция.

Попробуем подобрать параметр β_0 так, чтобы формула (15.41) была точной для многочленов второй степени, в частности, для функции $y(x) \equiv x^2$. Для этой функции $y' = f(x, y(x)) = 2x$, и через i-й узел и шаг h равенство (15.41)

записывается так:

$$(x_i+h)^2 = x_i^2 + h[\beta_0\, 2(x_i+h)+(1-\beta_0)2x_i].$$

После элементарных упрощений в последнем равенстве приходим к значению параметра $\beta_0=\dfrac{1}{2}$, подстановка которого в (15.41) приводит к методу

$$y_{i+1} = y_i + \frac{h}{2}(f_{i+1}+f_i).$$

В нем мы узнаём ***метод трапеций*** (14.15), единственный из семейства методов (15.41) (иначе, единственный из одношаговых методов семейства (15.35)), имеющий второй порядок точности.

Пусть теперь в общей формуле (15.35) $m=2$, т.е. будем рассматривать всевозможные **двухшаговые методы** этого семейства:

$$y_{i+1} = \alpha_1 y_i + \alpha_2 y_{i-1} + h(\beta_0 f_{i+1} + \beta_1 f_i + \beta_2 f_{i-1}). \qquad (15.42)$$

Потребуем, чтобы такой метод имел, по меньшей мере, второй порядок точности, т.е. чтобы подстановка в (15.42) вместо y_j значений $y(x_j)$ для $y(x)=x^2$ оставляла это равенство точным. Учтем, что оно должно быть верным при любом $i=1,2,3,\ldots$, и что, в силу однородности функции — предполагаемого решения $y(x)=x^2$, можно считать $i=1$ и, соответственно, $x_0=0$, $x_1=h$, $x_2=2h$. В этих точках решение $y=x^2$ принимает значения $0, h^2, 4h^2$, а его производная, т.е. функция $f(x,y(x))=2x$ — значения $0, 2h, 4h$. Следовательно, при $i=1$ равенство (15.42) приобретает вид

$$4h^2 = \alpha_1 h^2 + 4\beta_0 h^2 + 2\beta_1 h^2.$$

Записав для этого случая еще условия согласованности (15.37), (15.39), получаем систему трех линейных уравнений с пятью неизвестными параметрами:

$$\begin{cases} \alpha_1+\alpha_2=1, \\ \beta_0+\beta_1+\beta_2=\alpha_1+2\alpha_2, \\ \alpha_1+4\beta_0+2\beta_1=4. \end{cases} \qquad (15.43)$$

Будем считать свободными параметры

$$\alpha := \alpha_1 \quad \text{и} \quad \beta := \beta_0.$$

Тогда из системы (15.43) находим выражения остальных параметров:

$$\alpha_2 = 1-\alpha, \quad \beta_1 = 2-2\beta-\frac{\alpha}{2}, \quad \beta_2 = \beta-\frac{\alpha}{2}.$$

Их подстановка в (15.42) приводит к двухпараметрическому семейству двухшаговых методов, имеющих, по крайней мере, второй алгебраический порядок точности:

$$y_{i+1} = \alpha y_i + (1-\alpha)y_{i-1} + h\Big[\beta f_{i+1} + \left(2-2\beta-\frac{\alpha}{2}\right)f_i + \left(\beta-\frac{\alpha}{2}\right)f_{i-1}\Big]. \qquad (15.44)$$

Легко видеть, что при $\alpha=1$, $\beta=0$ это семейство содержит частным случаем метод Адамса–Башфорта второго порядка (см.(15.12)), а при $\alpha=1$, $\beta=\dfrac{1}{2}$ — метод Адамса–Моултона второго порядка (15.19) (он же — метод трапеций, единственный из содержащихся в (15.44) одношаговых методов).

Выберем еще одну степень свободы, потребовав, чтобы метод (15.44) имел третий порядок. Аналогично тому, как получали третье уравнение системы (15.43), полагаем $y(x)\equiv x^3$ и используем значения этой функции $0, h^3, 8h^3$ и значения ее производной $0, 3h^2, 12h^2$ на сетке $x_0=0$, $x_1=h$, $x_2=2h$ в равенстве (15.44) с фиксированным $i=1$. Имеем:

$$8h^3 = \alpha h^3 + 12\beta h^3 + \left(6-6\beta-\frac{3\alpha}{2}\right)h^3,$$

откуда получаем выражение $\beta=\dfrac{1}{3}+\dfrac{\alpha}{12}$. Таким образом, двухпараметрическое семейство методов, по крайней мере, второго порядка (15.44) превращается в однопараметрическое семейство методов, по крайней мере, третьего порядка

$$y_{i+1} = \alpha y_i + (1-\alpha)y_{i-1} + h\left[\left(\frac{1}{3}+\frac{\alpha}{12}\right)f_{i+1} + \left(\frac{4}{3}-\frac{2\alpha}{3}\right)f_i + \left(\frac{1}{3}-\frac{5\alpha}{12}\right)f_{i-1}\right]. \qquad (15.45)$$

Распоряжаться оставшимся единственным параметром α в (15.45) можно из разных соображений. Например, чтобы формула (15.45) была явной, нужно обнулить коэффициент при f_{i+1}, т.е. положить $\alpha=-4$. В результате этого приходим к двухшаго-

вому методу третьего порядка (не встречавшемуся ранее)

$$y_{i+1} = 5y_{i-1} - 4y_i + h(4f_i + 2f_{i-1}). \qquad (15.46)$$

Чтобы метод (15.45) относился к семейству (15.34) методов Адамса, следует обнулить коэффициент при y_{i-1}, т.е. взять $\alpha = 1$. При этом получаем известный метод Адамса–Моултона третьего порядка (15.20). Легко подметить, что при $\alpha = 0$ формула (15.45) определяет метод Симпсона, иначе, вторую формулу Милна (15.31), имеющую, как известно, четвертый порядок точности (см. выражение ее остаточного члена, точнее, локальной ошибки в формуле (15.30)).

Как будет выяснено позже, свободный параметр в многошаговых методах может потребоваться для того, чтобы за его счет улучшить численную устойчивость метода.

Замечание 15.2. Существуют линейные многошаговые методы, не вписывающиеся в общую формулу (15.35). К таким относятся так называемые ***методы с забеганием вперед.*** В основе вывода этих методов тоже лежит интегро-интерполяционный подход, но, в отличие от методов Адамса, опирающихся на второй интерполяционный многочлен Ньютона при аппроксимации подынтегральной функции f в интегральном равенстве (15.3), здесь в том же равенстве (15.3) применяется интерполяционный многочлен Бесселя (8.33). Он использует более близкую к промежутку $[x_i, x_{i+1}]$ информацию о функции, что обеспечивает большую локальную точность. Однако в результате получаются формулы, в которых участвуют недоступные для непосредственного (и даже неявного) вычисления значения, следующие за y_{i+1}. Например, основная формула наиболее известного представителя методов с забеганием вперед — ***метода Коуэлла четвертого порядка*** — имеет вид

$$y_{i+1} = y_i + \frac{h}{24}[-f_{i+2} + 13f_{i+1} + 13f_i - f_{i-1}]. \qquad (15.47)$$

Ясно, что одиночное применение таких формул невозможно, и методы с забеганием вперед реализуют совместно с другими, например, с методами Адамса. Так, выполнение одного шага метода Коуэлла (15.47) может осуществляться по следующему алгоритму.

Считая уже известными с предыдущего шага значения y_{i-1}, $f_{i-1} := f(x_{i-1}, y_{i-1})$, y_i, вычисляем значение $f_i := f(x_i, y_i)$; после этого проводим последовательные сближения по формулам, где $f_{i+1}^{(j)} := f\left(x_{i+1}, y_{i+1}^{(j)}\right)$:

$$y_{i+1}^{(1)} = y_i + \frac{h}{2}(3f_i - f_{i-1})$$

(методом Адамса–Башфорта второго порядка),

$$y_{i+1}^{(2)} = y_i + \frac{h}{12}\left(5f_{i+1}^{(1)} + 8f_i - f_{i-1}\right)$$

(методом Адамса–Моултона третьего порядка),

$$y_{i+2} = y_{i+1}^{(2)} + \frac{h}{12}\left(23f_{i+1}^{(2)} - 16f_i + 5f_{i-1}\right)$$

(методом Адамса–Башфорта третьего порядка),

$$y_{i+1}^{(3)} = y_i + \frac{h}{24}\left(-f_{i+2} + 13f_{i+1}^{(2)} + 13f_i - f_{i-1}\right)$$

(методом Коуэлла четвертого порядка);

после проверки на точность полагаем $y_{i+1} := y_{i+1}^{(3)}$.

Более подробно методы с забеганием вперед рассмотрены, например, в [100].

15.5. О ЧИСЛЕННОМ РЕШЕНИИ СИСТЕМ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ ПЕРВОГО ПОРЯДКА

Пусть требуется найти решение задачи Коши для системы обыкновенных дифференциальных уравнений первого порядка, разрешенных относительно производных:

$$\begin{cases} y_1' = f_1(x, y_1, \ldots, y_n), \\ y_2' = f_2(x, y_1, \ldots, y_n), \\ \ldots\ldots\ldots\ldots \\ y_n' = f_n(x, y_1, \ldots, y_n), \end{cases} \qquad \begin{cases} y_1(x_0) = y_1^0, \\ y_2(x_0) = y_2^0, \\ \ldots\ldots \\ y_n(x_0) = y_n^0. \end{cases}$$

Введем следующие векторные обозначения:

$$Y := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad Y' := \begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_n' \end{pmatrix}, \quad F(x, Y) := \begin{pmatrix} f_1(x, y_1, \ldots, y_n) \\ f_2(x, y_1, \ldots, y_n) \\ \ldots\ldots\ldots \\ f_n(x, y_1, \ldots, y_n) \end{pmatrix}, \quad Y_0 := \begin{pmatrix} y_1^0 \\ y_2^0 \\ \vdots \\ y_n^0 \end{pmatrix}.$$

С ними данная задача Коши принимает вид

$$Y' = F(x, Y), \quad Y(x_0) = Y_0, \qquad (15.48)$$

по форме точно такой же, как и рассматриваемая до сих пор задача (15.1)–(15.2).

19 Б-223

Легко понять, что к векторному дифференциальному уравнению (15.48), в принципе, можно применить любой из численных методов, изучавшихся в этой и в предыдущей главах (благодаря линейной структуре методов).

При этом скалярными величинами в формулах, определяющих методы, являются только независимая переменная x и расчетный шаг h; всем остальным величинам соответствуют введенные выше векторы размерности n. Следует лишь учесть, что при контроле точности вместо модуля нужно использовать норму вектора (например, норму-максимум).

Пример 15.1. Дана начальная задача

$$\begin{cases} y' = e^{-y^2-z^2} + 2x, \\ z' = 2y^2 + z, \end{cases} \quad \begin{cases} y(0) = 0.5, \\ z(0) = 1 . \end{cases}$$

Сделаем три шага явно-неявным методом Эйлера (предиктор-корректорным методом Адамса первого порядка), полагая $h = 0.1$. Предварительно для размерности $n = 2$ выполним покомпонентную запись расчетных формул. Для данной системы они выглядят так:

$$\begin{cases} \tilde{y}_{i+1} = y_i + 0.1(e^{-y_i^2-z_i^2} + 2x_i), \\ \tilde{z}_{i+1} = z_i + 0.1(2y_i^2 + z_i); \end{cases}$$

$$\begin{cases} y_{i+1} = y_i + 0.1(e^{-\tilde{y}_{i+1}^2-\tilde{z}_{i+1}^2} + 2x_i + 0.2), \\ z_{i+1} = z_i + 0.1(2\tilde{y}_{i+1}^2 + \tilde{z}_{i+1}). \end{cases}$$

Результаты счета с небольшой точностью по этим формулам при $i = 0, 1, 2$ представлены следующей таблицей.

i	x_{i+1}	$\tilde{y}_{i+1}$	$\tilde{z}_{i+1}$	y_{i+1}	z_{i+1}	$\max\{\lvert y_{i+1} - \tilde{y}_{i+1}\rvert, \lvert z_{i+1} - \tilde{z}_{i+1}\rvert\}$
0	0.1	≈0.5287	1.15	≈0.5401	≈1.1709	≈0.02
1	0.2	≈0.5791	≈1.3463	≈0.5918	≈1.3726	≈0.03
2	0.3	≈0.6425	≈1.5799	≈0.6573	≈1.6132	≈0.03

15.6. ЧИСЛЕННОЕ РЕШЕНИЕ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ ВЫСШИХ ПОРЯДКОВ. МЕТОДЫ АДАМСА–ШТЁРМЕРА

Одним из основных способов численного решения начальных задач для дифференциальных уравнений высших порядков является сведение их к соответствующим задачам для систем уравнений первого порядка.

Будем рассматривать уравнение второго порядка

$$y'' = f(x, y, y'), \quad x \in [x_0, b] \tag{15.49}$$

с начальными условиями

$$y(x_0) = y_0, \quad y'(x_0) = y'_0. \tag{15.50}$$

Введя новую переменную равенством $z = y'$, от уравнения (15.49) переходим к эквивалентной ему системе

$$\begin{cases} y' = z , \\ z' = f(x, y, z) . \end{cases} \tag{15.51}$$

Начальные условия (15.50) для нее перепишутся в виде

$$\begin{cases} y(x_0) = y_0, \\ z(x_0) = y'_0 . \end{cases} \tag{15.52}$$

К задаче Коши (15.51)–(15.52), в соответствии со сказанным в предыдущем параграфе, можно применить любой численный процесс из рассмотренных.

Альтернативой способу сведения к системам могут послужить методы, выводимые специально для уравнений высших порядков, например, на интегро-интерполяционной основе. Покажем, как может быть построен такой метод типа метода Адамса для задачи (15.49)–(15.50).

Будем считать, что уже известны значения $y_i \approx y(x_i)$, $y'_i \approx y'(x_i)$ и несколько предыдущих приближенных значений решения $y(x)$ и его производной $y'(x)$. Требуется получить формулы для вычисления величин y_{i+1} и y'_{i+1} — приближенных значений функций $y(x)$ и $y'(x)$ в очередной расчетной точке $x_{i+1} = x_i + h$. Для их вывода проинтегрируем уравнение (15.49) на промежутке $[x_i, x]$, в результате чего приходим к интегральному уравнению

$$y'(x) = y'(x_i) + \int_{x_i}^{x} f(t, y, y')dt \tag{15.53}$$

относительно неизвестной функции $y'(x)$. Зафиксируем в нем $x = x_{i+1}$, заменим $y'(x_i)$ известным приближенным значением y'_i и подставим в интеграле $\int_{x_i}^{x_{i+1}} f(x, y(x), y'(x))dx$ вместо функции

$f(x, y(x), y'(x))$ интерполяционный многочлен Ньютона (второй)

$$P_k(x) = P_k(x_i + qh) = f_i + q\Delta f_{i-1} + \frac{q(q+1)}{2!}\Delta^2 f_{i-2} + \ldots$$
$$\ldots + \frac{q(q+1)\ldots(q+k-1)}{k!}\Delta^k f_{i-k}, \quad (15.54)$$

построенный по значениям $f_j := f(x_j, y_j, y'_j)$ (где $j = i-k, i-k+1, \ldots, i-1, i$). После интегрирования получаем конечноразностную формулу

$$y'_{i+1} = y'_i + h\left(f_i + \frac{1}{2}\Delta f_{i-1} + \frac{5}{12}\Delta^2 f_{i-2} + \right.$$
$$\left. + \frac{3}{8}\Delta^3 f_{i-3} + \frac{251}{720}\Delta^4 f_{i-4} + \ldots\right), \quad (15.55)$$

которая представляет собой эктраполяционный метод Адамса–Башфорта (см. (15.10)), применяемый здесь для нахождения приближенных значений производной, а не самого решения.

Для выражения значения $y_{i+1} \approx y(x_{i+1})$ решения данной задачи через те же конечные разности, что участвуют в формуле (15.55), проинтегрируем равенство (15.53) в пределах от x_i до x_{i+1}. Имеем точное равенство

$$y(x_{i+1}) = y(x_i) + \int_{x_i}^{x_{i+1}} [y'(x_i) + \int_{x_i}^{x} f(t, y, y')dt]\,dx,$$

которое заменяем приближенным равенством

$$y(x_{i+1}) \approx y_i + hy'_i + \int_{x_i}^{x_{i+1}} \int_{x_i}^{x} f(t, y, y')\,dt\,dx. \quad (15.56)$$

Результат подстановки в правую часть последнего вместо $f(t, y(t), y'(t))$ интерполяционного многочлена $P_k(t)$ вида (15.54) принимаем за искомое значение y_{i+1}, т.е. полагаем

$$y_{i+1} = y_i + hy'_i + \int_{x_i}^{x_{i+1}} \int_{x_i}^{x} P_k(t)\,dt\,dx. \quad (15.57)$$

Сделав замены переменных интегрирования $x = x_i + ph$ и $t = x_i + qh$, выполним двойное интегрирование многочлена:

$$\int_{x_i}^{x_{i+1}} \int_{x_i}^{x} P_k(t)\,dt\,dx = h^2 \int_0^1 dp \int_0^p P_k(x_i + qh)\,dq =$$
$$= h^2 \int_0^1 dp \int_0^p \left[f_i + q\Delta f_{i-1} + \frac{q(q+1)}{2!}\Delta^2 f_{i-2} + \frac{q(q+1)(q+2)}{3!}\Delta^3 f_{i-3} + \right.$$
$$\left. + \frac{q(q+1)(q+2)(q+3)}{4!}\Delta^4 f_{i-4} + \ldots \right] dq =$$
$$= h^2 \int_0^1 \left[f_i p + \frac{p^2}{2}\Delta f_{i-1} + \left(\frac{p^3}{6} + \frac{p^2}{4}\right)\Delta^2 f_{i-2} + \frac{1}{6}\left(\frac{p^4}{4} + p^3 + p^2\right)\Delta^3 f_{i-3} + \right.$$
$$\left. + \frac{1}{24}\left(\frac{p^5}{5} + \frac{3p^4}{2} + \frac{11p^3}{3} + 3p^2\right)\Delta^4 f_{i-4} + \ldots \right] dp =$$
$$= h^2 \left(\frac{1}{2} f_i + \frac{1}{6}\Delta f_{i-1} + \frac{1}{8}\Delta^2 f_{i-2} + \frac{19}{180}\Delta^3 f_{i-3} + \frac{3}{32}\Delta^4 f_{i-4} + \ldots \right).$$

Подставляя это выражение двойного интеграла в (15.57), получаем конечноразностную формулу для приближенного вычисления самого решения $y(x)$ в точке x_{i+1}:

$$y_{i+1} = y_i + hy'_i + \frac{h^2}{2}\left(f_i + \frac{1}{3}\Delta f_{i-1} + \frac{1}{4}\Delta^2 f_{i-2} + \right.$$
$$\left. + \frac{19}{90}\Delta^3 f_{i-3} + \frac{3}{16}\Delta^4 f_{i-4} + \ldots \right). \quad (15.58)$$

Совокупность формул (15.55) и (15.58) определяют для задачи (15.49)–(15.50) семейство многошаговых экстраполяционных методов Адамса в конечноразностном виде. Фиксируя в них порядок последней используемой разности, тем самым задаем шаговость и порядок метода*), причем первые из отбрасываемых слагаемых (с учетом множителя за скобками) грубо характеризуют шаговую погрешность вычисления y'_{i+1} и y_{i+1}.

Например, при $k = 3$ в (15.54), т.е. при использовании первых, вторых и третьих разностей, из (15.55) и (15.58) получаем для задачи (15.49)–(15.50) следующий ***явный четырехшаговый***

*) Если последними учитываются k-е разности, то в скобках формул (15.55) и (15.58) берется $k+1$ слагаемых, что как раз соответствует шаговости и порядку метода.

метод Адамса четвертого порядка (в ординатном виде):

$$\begin{cases} y'_{i+1} = y'_i + \dfrac{h}{24}(55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}), \\ y_{i+1} = y_i + hy'_i + \dfrac{h^2}{360}(323f_i - 264f_{i-1} + 159f_{i-2} - 38f_{i-3}). \end{cases} \quad (15.59)$$

Главная часть локальной ошибки первой из формул (15.59) составляет величину $\frac{251}{720}h\Delta^4 f_{i-4} = O(h^5)$, а второй — величину $\frac{3}{32}h^2\Delta^4 f_{i-4} = O(h^6)$.

Нетрудно вывести и неявные методы Адамса для задачи (15.49)–(15.50), подменяя функцию f в равенствах (15.53) и (15.56) интерполяционным многочленом вида (15.54) с увеличенными на единицу индексами, т.е. многочленом $P_k(x_{i+1} + qh)$. Совершенно очевидно, что для вычисления y'_{i+1} при этом будет использована выведенная ранее интерполяционная формула Адамса–Моултона (15.17), где вместо y_i, y_{i+1} следует записать соответственно y'_i, y'_{i+1}. Вывод неявной формулы для вычисления y_{i+1} предоставим читателю. На базе явных и неявных методов можно устраивать предиктор-корректорные алгоритмы с пошаговым контролем точности.

Рассмотрим, наконец, один частный вид уравнений (15.49), для которых вычисления по методу Адамса можно сделать более лаконичными. Именно, будем строить метод решения задачи Коши для уравнения

$$y'' = f(x, y), \quad (15.60)$$

правая часть которого не содержит производной. Обратив внимание на то, что в таком случае при вычислении значений f_j значения производной не требуются, вместо двух формул, реализующих явный метод Адамса (15.55), (15.58), можно попытаться ограничиться одной формулой (15.58), если удастся избавиться в ней от слагаемого hy'_i. С этой целью выведем вспомогательную формулу типа (15.58), также содержащую слагаемое hy'_i, что позволит исключить его алгебраическим сложением формул.

Проинтегрируем равенство (15.53), в котором вместо функции $f(x, y, y')$ будем подразумевать функцию $f(x, y)$ — правую часть уравнения (15.60), на промежутке $[x_{i-1}, x_i]$. Имеем

$$y(x_i) = y(x_{i-1}) + hy'(x_i) + \int_{x_{i-1}}^{x_i}\int_{x_i}^{x} f(t, y(t))\,dt\,dx. \quad (15.61)$$

Подставляя сюда y_{i-1} вместо $y(x_{i-1})$, y'_i вместо $y'(x_i)$ и тот же многочлен Ньютона (15.54) вместо подынтегральной функции f, получим формулу, аналогичную (15.57):

$$y_i = y_{i-1} + hy'_i + \int_{x_{i-1}}^{x_i}\int_{x_i}^{x} P_k(t)\,dt\,dx\ . \quad (15.62)$$

С теми же заменами переменных $x = x_i + ph$ и $t = x_i + qh$ производим двойное интегрирование в (15.62) (см. выкладки при выводе формулы (15.58)):

$$\int_{x_{i-1}}^{x_i}\int_{x_i}^{x} P_k(t)\,dt\,dx = h^2\int_{-1}^{0}dp\int_{0}^{p} P_k(x_i + qh)dq =$$

$$= h^2\left(-\frac{1}{2}f_i + \frac{1}{6}\Delta f_{i-1} + \frac{1}{24}\Delta^2 f_{i-2} + \frac{1}{45}\Delta^3 f_{i-3} + \frac{7}{480}\Delta^4 f_{i-4} + \ldots\right).$$

Будем теперь рассматривать совместно получающуюся подстановкой в (15.62) этого выражения двойного интеграла формулу

$$y_i = y_{i-1} + h\,y'_i + \frac{h^2}{2}\left(-f_i + \frac{1}{3}\Delta f_{i-1} + \frac{1}{12}\Delta^2 f_{i-2} + \right.$$

$$\left. + \frac{2}{45}\Delta^3 f_{i-3} + \frac{7}{240}\Delta^4 f_{i-4} + \ldots\right) \quad (15.63)$$

и формулу (15.58), где, как и в (15.63), считаем $f_j := f(x_j, y_j)$. Вычитая из равенства (15.58) равенство (15.63), приходим к формуле

$$y_{i+1} = 2y_i - y_{i-1} + h^2\left(f_i + \frac{1}{12}\Delta^2 f_{i-2} + \right.$$

$$\left. + \frac{1}{12}\Delta^3 f_{i-3} + \frac{19}{240}\Delta^4 f_{i-4} + \ldots\right), \quad (15.64)$$

которая называется ***формулой Штёрмера****) или ***Адамса–Штёрмера***. Как видим, ***метод Штёрмера*** решения начальной зада-

*) Штёрмер (Стёрмер) Фредрик Карл Мюлерц (1874–1957) — норвежский геофизик и математик.

чи (15.60), (15.50) определяется одной явной формулой (15.64) и является более простым и экономичным по сравнению с базовым для него методом Адамса.

Так же просто выводится и неявная формула Штёрмера, которая может быть привлечена к организации вычислений по схеме «предсказание-уточнение».

Замечание 15.3. Многошаговые методы прогноза и коррекции высоких порядков позволяют строить алгоритмы с автоматическим выбором шага, отличающиеся априорно высокой эффективностью. Однако практика использования таких простых алгоритмов показывает, что их реальная эффективность значительно ниже предполагаемой из-за частых изменений величины шага. Поэтому непосредственно методы Адамса прогноза и коррекции на промежутках значительной протяженности целесообразно применять в тех случаях, когда вычисления можно проводить с некоторым наперед известным шагом. В противном случае применяют более эффективные, создаваемые на той же многошаговой основе, специальные алгоритмы Гира [6, 188].

УПРАЖНЕНИЯ

15.1. Запишите явные и неявные формулы Адамса пятого порядка точности (в ординатном виде).

15.2. Получите приближенные критерии пошагового контроля точности в предиктор-корректорных методах Адамса:

а) первого порядка;
б) второго порядка;
в) третьего порядка.

15.3. Интегро-интерполяционным способом выведите предиктор-корректорные формулы третьего порядка точности, используя интегральное равенство типа (15.3) на промежутке $[x_{i-2}, x_{i+1}]$.

Сравните пошаговую точность полученного метода с аналогичным показателем предиктор-корректорного метода Адамса третьего порядка (см. упр. 15.2, в).

15.4. Выведите вторую формулу Милна и главный член ее локальной ошибки, интегрируя уравнение $y' = f(x, y)$ на промежутке $[x_{i-1}, x_{i+1}]$ и приближая подынтегральную функцию $f(x, y(x))$ первым многочленом Ньютона третьей степени с базовым узлом x_{i-1}.

15.5. Проверьте выполнение условий согласованности параметров в методах Адамса, Милна и в методе, выведенном в упр.15.3.

15.6. Исходя из общего вида линейных многошаговых методов и понятия алгебраического порядка точности, опишите различные семейства трехшаговых методов разных порядков.

15.7. Для задачи, рассматривавшейся в примере 14.1 § 14.3, выполните один полный шаг метода Коуэлла (15.47), считая $h = 0.1$. Окончательный результат $y_2 \approx y(0.2)$ и все промежуточные результаты последовательных сближений сравните с точными значениями решения в соответствующих точках.

15.8. А) Убедитесь в справедливости формулы Коуэлла (15.47).

Б) Постройте метод не ниже третьего порядка точности на основании интерполяционной формулы Стирлинга (8.32).

15.9. Запишите расчетные формулы получения приближенных значений решения задачи Коши для двумерной системы ОДУ предиктор-корректорным методом Адамса второго порядка. Выполните по этим формулам расчеты для задачи, поставленной в примере 15.1 § 15.5.

15.10. А) Выведите неявную конечноразностную формулу Адамса для численного решения задачи

$$y'' = f(x, y, y'), \quad y(x_0) = y_0, \quad y'(x_0) = y'_0.$$

Б) Запишите четырехшаговый предиктор-корректорный метод Адамса для этой задачи и выведите приближенное правило пошагового контроля точности.

15.11. Запишите частные формулы Штёрмера в ординатном виде:

а) двухшаговые;
б) трехшаговые;
в) четырехшаговые.

15.12. А) Выведите неявную конечноразностную формулу Штёрмера, рассмотрите ее частные случаи.

Б) Запишите четырехшаговый предиктор-корректорный метод Адамса–Штёрмера и получите формулу для пошагового контроля точности этого метода.

ГЛАВА 16 О ПРОБЛЕМАХ ЧИСЛЕННОЙ УСТОЙЧИВОСТИ

Описывается общий принцип построения методов решения задач численного анализа и определяются связанные с этим понятия аппроксимации, устойчивости, сходимости. Показывается реализация этого принципа при аппроксимации разностными уравнениями начальных задач для обыкновенных дифференциальных уравнений. Приводятся краткие сведения о решениях линейных разностных уравнений с постоянными коэффициентами, которые затем используются при изучении поведения решений и ошибок некоторых простейших схем. Вводится определение устойчивости по Дальквисту, дающее возможность легко выявлять неустойчивые многошаговые методы. С помощью примеров формируется представление о жестких начальных задачах, дается определение жесткой системы. Через понятие области устойчивости определяется A*-устойчивость и* $A(\alpha)$*-устойчивость разностных методов, обеспечивающие приемлемость этих методов для решения жестких задач (в частности, выделяются чисто неявные методы дифференцирования назад, обладающие таким свойством).*

16.1. ОБЩАЯ СХЕМА РЕШЕНИЯ ЗАДАЧ ЧИСЛЕННОГО АНАЛИЗА. АППРОКСИМАЦИЯ, УСТОЙЧИВОСТЬ, СХОДИМОСТЬ

Большинство задач численного анализа, будь то начальные или граничные задачи для дифференциальных уравнений, интегральные уравнения и т.п., достаточно естественно можно записать в виде уравнения

$$F(y) = z, \tag{16.1}$$

где $F: Y \to Z$ — линейный или нелинейный оператор (или функционал), переводящий элементы метрического пространства Y в метрическое пространство Z. Суть приближенных методов решения таких задач, как ясно видно из всего предшествующего материала, заключается в том, что уравнение (16.1) заменяется близким ему, в некотором смысле более простым (обычно конечномерным) уравнением

$$F_n(y_n) = z_n. \tag{16.2}$$

Это уравнение определяется оператором $F_n: Y_n \to Z_n$, соответствующим данному оператору F и действующим из метрического

пространства Y_n в метрическое пространство Z_n. При этом элементы $y_n \in Y_n$ и $z_n \in Z_n$ рассматриваются как образы элементов $y \in Y$ и $z \in Z$ соответствующих исходных пространств, и такая связь задается некоторыми ***операторами сноса*** $\varphi_n: Y \to Y_n$ и $\psi_n: Z \to Z_n$, т.е. равенствами

$$y_n = \varphi_n(y), \quad z_n = \psi_n(z).$$

Обратное соответствие между пространствами Y_n и Y, а точнее, между Y_n и некоторым подпространством пространства Y, устанавливается с помощью ***оператора восполнения*** φ_n^{-1}:

$$y = \varphi_n^{-1}(y_n). \tag{16.3}$$

Схематично связь между четырьмя фигурирующими здесь пространствами показана на рис.16.1а.

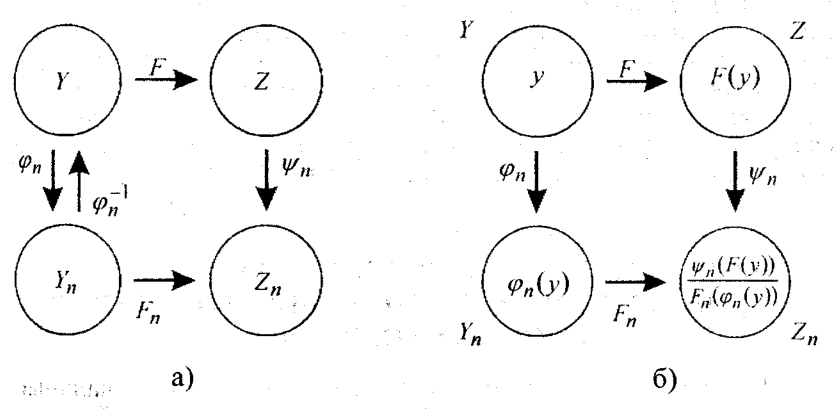


Рис. 16.1. *К связи между задачами* (16.1) *и* (16.2)

Чтобы понять, как можно описать близость задач (16.1) и (16.2), проследим связь между пространствами Y, Z, Z_n и Y_n на уровне элементов (рис.16.1б).

Зафиксируем некоторый элемент $y \in Y$. Его образом в пространстве Z, благодаря данному оператору F, будет элемент $z = F(y)$, а в пространстве Z_n с помощью оператора сноса ψ_n получаем элемент $\psi_n(F(y))$. С другой стороны, тому же элементу y оператор сноса φ_n ставит в соответствие элемент $y_n = \varphi_n(y)$ пространства Y_n, а ему, в свою очередь, новый оператор F_n сопоставляет элемент $F_n(\varphi_n(y))$ пространства Z_n. Так как элементы $\tilde{z}_n := \psi_n(F(y))$ и $\hat{z}_n := F_n(\varphi_n(y))$ служат образами одного и того же элемента y из Y в одном и том же пространст-

ве Z_n, то по близости между ними можно судить о том, насколько близки операторы F_n и F.

Определение 16.1. *Говорят, что уравнение*

$$F_n(y_n)=\psi_n(z) \qquad (16.4)$$

аппроксимирует уравнение (16.1) (*оператор* F_n *аппроксимирует оператор* F), *если для любых* y *из* $D(F)\subseteq Y$ ***мера аппроксимации***

$$\rho_{Z_n}(F_n(\varphi_n(y)),\ \psi_n(F(y))) \qquad (16.5)$$

стремится к нулю при $n\to\infty$.

(Здесь $\rho_{Z_n}(\cdot,\cdot)$ обозначает метрику, т.е. расстояние между указанными в скобках элементами пространства Z_n).

Чтобы иметь возможность грубо сравнивать качество различных моделей типа (16.4) задачи (16.1), часто используют понятие ***порядка аппроксимации***, связывая стремление к нулю меры аппроксимации (16.5) с порядком убывания какой-либо зависящей от n малой величины (***шага аппроксимации***).

Предположим, что решения $y^*\in Y$ и $y_n^*\in Y_n$ уравнений соответственно (16.1) и (16.4) существуют и единственны. Поскольку решение задачи (16.1) ищется в пространстве Y, ее ***приближенным решением*** считается получаемый с помощью оператора восполнения (16.3) элемент

$$y^{(n)}=\varphi_n^{-1}\left(y_n^*\right),$$

решение же y_n^* задачи (16.4) называется ***каркасом приближенного решения****). Наверное, главным вопросом любой теории приближенных методов решения задач вида (16.1) является вопрос о том, можно ли приближенным решением $y^{(n)}$ сколь угодно хорошо отразить поведение точного решения y^*, иначе, вопрос о сходимости $y^{(n)}$ к y^*.

Определение 16.2. *Говорят, что имеет место* ***сходимость приближенных решений*** $y^{(n)}$ *к точному решению* y^* *уравнения* (16.1), *если* $\rho_Y\left(y^*,\ y^{(n)}\right)\xrightarrow[n\to\infty]{}0$ (т.е. если расстояние между $y^{(n)}$ и y^* по метрике пространства Y может быть сделано сколь угодно малым).

*) Типичная ситуация: вместо точного решения — функции — находят таблицу ее приближенных значений (каркас), а затем производят восполнение, например, посредством сплайн-интерполирования; естественно, что получаемое при этом приближенное решение принадлежит пространству более узкому, чем исходное пространство решений. Не следует здесь воспринимать φ_n^{-1} как обозначение обратного (в классическом смысле) к φ_n оператора.

Очевидно, имеет смысл рассматривать также ***сходимость каркасов приближенных решений***, понимая под этим выполнение условия

$$\rho_{Y_n}\left(y_n^*,\ \varphi_n\left(y^*\right)\right)\xrightarrow[n\to\infty]{}0.$$

Подобно порядку аппроксимации, вводится понятие порядка сходимости приближенных решений и (или) их каркасов.

Наличие фактических оценок величин $\rho_{Y_n}\left(y_n^*,\varphi_n\left(y^*\right)\right)$ позволяет не только делать выводы о сходимости приближенных решений и (или) их каркасов, но и указывать погрешности получаемых приближений к решению.

Вопрос о сходимости приближенных решений $y^{(n)}$ к y^* тесно связан с тем, можем ли мы надежно получать каркас решения y_n^*, решая упрощенную задачу (16.4) (в предположении, что последняя аппроксимирует данную задачу (16.1) в смысле определения 16.1)? Дело в том, что эта упрощенная задача (16.4), вообще говоря, другая и тоже решается приближенно; улучшение качества аппроксимации, т.е. уменьшение ее меры (16.5), влечет увеличение размерности n решаемой задачи (16.4), а значит, рост объема вычислений, что, в свою очередь, может привести к катастрофическому нарастанию вычислительных погрешностей.

Определение 16.3. *Вычислительный процесс называется устойчивым, если малые погрешности исходных данных вызывают малые погрешности результата* (рост погрешностей ограничен).

В определении 16.3 заложено скорее понятие, чем строгое определение численной (иначе, вычислительной) устойчивости. Имеется ряд более конкретных определений численной устойчивости применительно к более конкретно поставленным задачам приближенных вычислений (см., например, далее определение 16.4 в § 16.5, определения 16.8 и 16.9 в § 16.7).

Одна из теорем численного анализа гласит: «*Аппроксимация плюс устойчивость влечет сходимость*». Ее доказательство (и более корректные формулировки) можно найти в книгах [12, 67, 92, 100, 152, 158 и др.].

16.2. ПРОСТЕЙШИЕ РАЗНОСТНЫЕ АППРОКСИМАЦИИ ЗАДАЧИ КОШИ. ГЛОБАЛЬНАЯ ПОГРЕШНОСТЬ МЕТОДА ЭЙЛЕРА

Вернемся к изучению численных процессов решения начальной задачи

$$\begin{aligned} &y' = f(x, y), \quad x \in [x_0, b], \\ &y(x_0) = y_0. \end{aligned} \tag{16.6}$$

Положив $h = \dfrac{b - x_0}{n}$, введем на отрезке $[x_0, b]$ равномерную ***сетку****)

$$\omega_h := \{ x_i \mid x_i = x_0 + ih \ (i = 0, 1, \ldots, n)\}. \tag{16.7}$$

Функции, определенные во всех ***узлах*** x_i ***сетки*** ω_h, называют ***сеточными функциями***. Например, если некоторая функция $y = \varphi(x)$ определена в узлах сетки ω_h, то сеточной функцией следует считать функцию $y_i = \varphi(x_i)$ дискретного аргумента $i = 0, 1, \ldots, n$, т.е. $(n+1)$-мерный вектор $(\varphi(x_0), \varphi(x_1), \ldots, \varphi(x_n))$. В свете сказанного в предыдущем параграфе бесконечномерная дифференциальная задача (16.6), состоящая в нахождении удовлетворяющей ей функции — решения $y = y(x)$, сводится к конечномерной задаче вычисления $(n+1)$-мерного вектора $(y(x_0), y(x_1), \ldots, y(x_n))$ — соответствующей этому решению $y(x)$ сеточной функции $y(x_i)$ на ω_h, другими словами, каркаса решения $y_0 \approx y(x_0)$, $y_1 \approx y(x_1)$, $\ldots$, $y_n \approx y(x_n)$. Одним из основных приемов такого сведения является уже использовавшийся в § 7.2 разностный подход, т.е. рассмотрение уравнения (16.6) в узлах сетки ω_h и аппроксимация в них производной y' через соседние значения сеточной функции, соответствующей исходному решению $y = y(x)$ (иначе, замена дифференциального оператора разностным).

Знание порядков используемых формул аппроксимации производной и их остаточных членов позволяет получать представление о порядке аппроксимации данной бесконечномерной задачи конечномерной, выводить оценки погрешностей приближенных решений (на сетке) и изучать устойчивость и сходимость каркасов решений. В связи с принятием за основу разностного подхода к построению конечномерных моделей дифференциальных уравнений, эти модели называют ***разностными уравнениями*** или ***разностными схемами.***

---

*) Естественно, сетки бывают и неравномерными.

Положим в (16.6) $x = x_i$ и к левой части полученного равенства

$$y'(x_i) = f(x_i, y(x_i))$$

применим простейшие аппроксимации первого (6.15), (6.14) и второго (6.19), (6.26), (6.27) порядков. Имеем:

$$\frac{y(x_{i+1}) - y(x_i)}{h} = f(x_i, y(x_i)) + O(h);$$

$$\frac{y(x_i) - y(x_{i-1})}{h} = f(x_i, y(x_i)) + O(h); \tag{16.8}$$

$$\frac{y(x_{i+1}) - y(x_{i-1})}{2h} = f(x_i, y(x_i)) + O(h^2); \tag{16.9}$$

$$\frac{-3y(x_i) + 4y(x_{i+1}) - y(x_{i+2})}{2h} = f(x_i, y(x_i)) + O(h^2);$$

$$\frac{y(x_{i-2}) - 4y(x_{i-1}) + 3y(x_i)}{2h} = f(x_i, y(x_i)) + O(h^2). \tag{16.10}$$

Отбрасывание в этих равенствах последнего слагаемого, характеризующего порядок аппроксимации, и замена в них точных значений $y(x_j)$ решения $y(x)$ в j-х узлах сетки ω_h приближенными значениями y_j приводит к следующим разностным схемам:

$$y_{i+1} = y_i + hf(x_i, y_i), \qquad i = 0, 1, \ldots, n-1; \tag{16.11}$$

$$y_{i+1} = y_i + hf(x_{i+1}, y_{i+1}), \qquad i = 0, 1, \ldots, n-1; \tag{16.12}$$

$$y_{i+1} = y_{i-1} + 2hf(x_i, y_i), \qquad i = 1, 2, \ldots, n-1; \tag{16.13}$$

$$y_{i+1} = 4y_i - 3y_{i-1} - 2hf(x_{i-1}, y_{i-1}), \qquad i = 1, 2, \ldots, n-1; \tag{16.14}$$

$$y_{i+1} = \frac{4}{3}y_i - \frac{1}{3}y_{i-1} + \frac{2h}{3}f(x_{i+1}, y_{i+1}), \quad i = 1, 2, \ldots, n-1. \tag{16.15}$$

Первые три из схем (16.11)–(16.15) определяют хорошо знакомые методы Эйлера: явный (14.8), неявный (14.14) и уточненный (14.19). Последние две схемы второго порядка представляют собой явный и неявный двухшаговые методы второго порядка, внимание которым будет уделено позже.

До сих пор нами изучались лишь локальные ошибки методов, т.е. ошибки, возникающие на одном текущем шаге в предположении, что исходным материалом для получения результата этого шага служат точные значения. Попытаемся теперь хотя бы

в простейших случаях изучить поведение ошибок в их взаимосвязи на соседних шагах и получить представление о процессе накопления методической погрешности к n-му шагу. Знание того, как изменяются погрешности от шага к шагу, позволяет делать выводы о численной устойчивости разностного метода и о сходимости поставляемых им каркасов приближенных решений, получать оценки глобальной погрешности.

Обозначим

$$\delta_i := y(x_i) - y_i \tag{16.16}$$

— разность между значением точного решения $y = y(x)$ задачи (16.6) в i-м узле сетки (16.7) и соответствующей компонентой каркаса решения, получаемого тем или иным разностным методом.

Предположим, что правая часть и решение данной задачи (16.6) обладает достаточной гладкостью и выполняются условия

$$\exists\ C_1, C_2 > 0 : \left|f'_y(x, y)\right| \le C_1, \quad |y''(x)| \le C_2 \quad \forall\, x \in [x_0, b]. \tag{16.17}$$

В этом предположении оценим глобальную погрешность метода Эйлера.

Вычитая из линеаризованного по формуле Тейлора выражения решения $y = y(x)$ в точке x_{i+1} получаемое явным методом Эйлера (16.11) значение y_{i+1}, в соответствии с обозначением (16.16) имеем:

$$\delta_{i+1} = y(x_{i+1}) - y_{i+1} =$$

$$= y(x_i) + y'(x_i)h + \frac{1}{2}y''(\Theta_i)h^2 - y_i - hf(x_i, y_i) =$$

$$= \delta_i + h[f(x_i, y(x_i)) - f(x_i, y_i)] + \frac{1}{2}y''(\Theta_i)h^2.$$

Применив к разности функций в последнем выражении формулу Лагранжа по второму аргументу, получаем

$$\delta_{i+1} = \delta_i + hf'_y(x_i, v_i)\delta_i + \frac{1}{2}y''(\Theta_i)h^2. \tag{16.18}$$

Таким образом, связь между ошибками в $(i+1)$-м и i-м узлах описывается разностным уравнением

$$\delta_{i+1} = A_i\delta_i + B_i, \quad i = 0, 1, \ldots, n-1, \tag{16.19}$$

где

$$A_i := 1 + hf'_y(x_i, v_i), \quad B_i := \frac{1}{2}y''(\Theta_i)h^2 \tag{16.20}$$

(Θ_i и v_i — некоторые точки из области задания и области значений решения $y(x)$ соответственно). Анализировать это уравнение будем позже, а здесь произведем оценивание абсолютной погрешности, используя условия (16.17). Благодаря им из (16.18) получаем рекуррентное неравенство

$$|\delta_{i+1}| \le A|\delta_i| + B, \quad i = 0, 1, \ldots, n-1,$$

где

$$A := 1 + C_1 h, \quad B := \frac{1}{2}C_2 h^2. \tag{16.21}$$

Итерирование этого неравенства дает

$$|\delta_{i+1}| \le A(A|\delta_{i-1}| + B) + B = A^2|\delta_{i-1}| + (A+1)B \le$$

$$\le A^2(A|\delta_{i-2}| + B) + (A+1)B = A^3|\delta_{i-2}| + \left(A^2 + A + 1\right)B \le \ldots$$

$$\ldots \le A^{i+1}|\delta_0| + \left(A^i + A^{i-1} + \ldots + A + 1\right)B = \frac{A^{i+1} - 1}{A - 1}B,$$

поскольку $\delta_0 = y(x_0) - y_0 = 0$, согласно начальному условию. При $i = n-1$ в соответствии с обозначением (16.21) отсюда получаем **оценку глобальной погрешности метода Эйлера**

$$|\delta_n| \le \frac{(1 + C_1 h)^n - 1}{C_1 h} \cdot \frac{1}{2}C_2 h^2 = \frac{C_2 h}{2C_1}\left[(1 + C_1 h)^n - 1\right].$$

Чтобы проще было судить о порядке глобальной погрешности, применим в правой части ее оценки формулу Ньютона n-й степени бинома. В результате имеем

$$|y(b) - y_n| \le \frac{C_2 h}{2C_1}(nC_1 h + o(h)) = \frac{1}{2}C_2(b - x_0)h + o\left(h^2\right) = O(h).$$

Как видим, *глобальная погрешность метода Эйлера имеет первый порядок относительно шага* h *и совпадает по порядку с погрешностью аппроксимации дифференциальной начальной задачи* (16.6) *дискретной задачей* (16.11) *с начальным значением* y_0.

Если при выводе разностных уравнений типа (16.19) для ошибок δ_i ограничиваться указанием только порядка свободных членов, то такие уравнения проще получать сравнением исследуемых разностных схем с приведшими к ним результатами аппроксимаций дифференциального оператора разностным. Продемонстрируем это на неявном (16.12) и уточненном (16.13) методах Эйлера.

Перепишем равенство (16.8) в виде

$$y(x_{i+1}) = y(x_i) + hf(x_{i+1}, y(x_{i+1})) + O(h^2)$$

и вычтем из него (16.12). Используя обозначение (16.16) и формулу конечных приращений Лагранжа, имеем

$$\delta_{i+1} = \delta_i + h[f(x_{i+1}, y(x_{i+1})) - f(x_{i+1}, y_{i+1})] + O(h^2) =$$

$$= \delta_i + hf'_y(x_{i+1}, v_{i+1})\delta_{i+1} + O(h^2),$$

откуда после разрешения полученного равенства относительно δ_{i+1} приходим к тому же разностному уравнению (16.19), в котором

$$A_i := \frac{1}{1 - hf'_y(x_{i+1}, v_{i+1})}, \quad B_i := \frac{O(h^2)}{1 - hf'_y(x_{i+1}, v_{i+1})}. \qquad (16.22)$$

Аналогично, сравнение равенства (16.13), определяющего уточненный метод Эйлера, с эквивалентным (16.9) равенством

$$y(x_{i+1}) = y(x_{i-1}) + 2hf(x_i, y(x_i)) + O(h^3)$$

дает

$$\delta_{i+1} = \delta_{i-1} + 2h[f(x_i, y(x_i)) - f(x_i, y_i)] + O(h^3) =$$

$$= \delta_{i-1} + 2hf'_y(x_i, v_i)\delta_i + O(h^3).$$

Следовательно, ошибка двухшагового метода (16.13), основанного на симметричной формуле второго порядка аппроксимации производной, удовлетворяет трехточечному рекуррентному соотношению

$$\delta_{i+1} = A_i\delta_i + \delta_{i-1} + O(h^3), \qquad (16.23)$$

где $A_i := 2hf'_y(x_i, v_i)$.

16.3. КРАТКИЕ СВЕДЕНИЯ О РЕШЕНИЯХ ЛИНЕЙНЫХ РАЗНОСТНЫХ УРАВНЕНИЙ С ПОСТОЯННЫМИ КОЭФФИЦИЕНТАМИ

Как видно из предыдущего параграфа и из предыдущей главы, численное решение дифференциального уравнения сводится к решению уравнения разностного, связывающего приближенные значения исходного решения в нескольких соседних узлах сетки, количество которых (минус один) определяет, сколькишаговым является численный метод. Разностным уравнениям удовлетворяют и ошибки методов в узлах (см., например, (16.19), (16.23)). Поэтому для анализа приближенных решений дифференциальных уравнений и их погрешностей важно иметь представление о решениях разностных уравнений.

Ограничимся некоторыми первичными сведениями о решениях ***линейных разностных уравнений m-го порядка с постоянными коэффициентами***, имеющих вид

$$u_{i+1} = a_1u_i + a_2u_{i-1} + \ldots + a_mu_{i-m+1} + b. \qquad (16.24)$$

Изучение таких уравнений проводится аналогично изучению линейных дифференциальных уравнений с постоянными коэффициентами. Так же, как и в дифференциальном случае, рассматривается соответствующее (16.24) ***однородное разностное уравнение***

$$U_{i+1} = a_1U_i + a_2U_{i-1} + \ldots + a_mU_{i-m+1}, \qquad (16.25)$$

ищется его ***общее решение*** U_i, представляющее собой линейную комбинацию m ***фундаментальных решений*** U_{ij} $(j = 1, 2, \ldots, m)$, находится какое-либо ***частное решение*** $\overline{u}_i$ неоднородного уравнения (16.24); тогда общее решение u_i уравнения (16.24) представляется суммой U_i и $\overline{u}_i$.

Подобно тому, как в случае линейного дифференциального уравнения решение соответствующего однородного ищут в виде показательной функции, здесь нетривиальное решение уравнения (16.25) естественно искать в виде

$$U_i = \lambda^i \qquad (16.26)$$

с некоторой неизвестной постоянной λ $(\neq 0)$. Подставляя (16.26) в (16.25), имеем

$$\lambda^{i+1} = a_1\lambda^i + a_2\lambda^{i-1} + \ldots + a_m\lambda^{i-m+1},$$

откуда после деления на λ^{i-m+1} приходим к уравнению

$$\lambda^m = a_1\lambda^{m-1} + a_2\lambda^{m-2} + \ldots + a_{m-1}\lambda + a_m. \qquad (16.27)$$

Это уравнение называют ***характеристическим*** по отношению к уравнению (16.25) (а значит, и к (16.24)), поскольку уравнение (16.25) имеет решения вида (16.26) только в том случае, если λ есть корень алгебраического уравнения (16.27).

Предположим, что все корни $\lambda_1, \lambda_2, \ldots, \lambda_m$ характеристического уравнения (16.27) — действительные различные. Тогда выражения

$$U_{i1} = \lambda_1^i, \quad U_{i2} = \lambda_2^i, \quad \ldots, \quad U_{im} = \lambda_m^i$$

образуют полную ***систему фундаментальных решений***, и общее решение этого уравнения есть

$$U_i = c_1\lambda_1^i + c_2\lambda_2^i + \ldots + c_m\lambda_m^i, \tag{16.28}$$

где $c_1, c_2, \ldots, c_m$ — произвольные постоянные.

Такой же вид (16.28) будет иметь общее решение уравнения (16.25) и в случае, когда среди корней уравнения (16.27) есть комплексные, но нет кратных. Если же некоторое число λ_j является k-кратным корнем характеристического уравнения, то, опять-таки, подобно дифференциальному случаю, ему будет соответствовать k фундаментальных решений [158]:

$$\lambda_j^i, \quad i\lambda_j^i, \quad \ldots, \quad i^{k-1}\lambda_j^i.$$

Обращаясь теперь к исходному неоднородному уравнению (16.24), непосредственной проверкой убеждаемся, что выражение

$$\overline{u}_i = \frac{b}{1 - a_1 - a_2 - \ldots - a_m} \tag{16.29}$$

можно считать его частным решением (если в (16.29) знаменатель не обращается в нуль). Таким образом, общее решение разностного уравнения (16.24) есть функция целочисленного аргумента i, имеющая вид

$$u_i = c_1\lambda_1^i + c_2\lambda_2^i + \ldots + c_m\lambda_m^i + \frac{b}{1 - a_1 - a_2 - \ldots - a_m}, \tag{16.30}$$

если λ_j $(j = 1, \ldots, m)$ — простые корни характеристического уравнения (16.27), и содержащая слагаемые вида

$$c_j\lambda_j^i + c_{j+1}i\lambda_j^i + \ldots + c_{j+k-1}i^{k-1}\lambda_j^i,$$

соответствующие каждому k-кратному корню λ_j.

Для разностных уравнений, как и для дифференциальных, также можно ставить начальные и краевые задачи, задавая значения u_i при определенных значениях i, что позволяет из общих решений фиксированием произвольных постоянных выделять частные решения, удовлетворяющие конкретной задаче. В этой главе для нас будут представлять интерес начальные задачи для $(m+1)$-точечных разностных уравнений (16.24), которые можно считать некоторыми m-шаговыми аппроксимациями задачи Коши (16.6).

16.4. УСТОЙЧИВОСТЬ И НЕУСТОЙЧИВОСТЬ НЕКОТОРЫХ ПРОСТЕЙШИХ РАЗНОСТНЫХ СХЕМ

Изучение устойчивости численных методов решения начальных задач (16.6) обычно проводят на простом уравнении вида

$$y' = py, \tag{16.31}$$

называемом в данном случае ***модельным уравнением***; будем пока считать здесь p вещественным параметром*). Его общее решение есть

$$y = Ce^{px},$$

и решение соответствующей ему задачи Коши с начальным условием $y(x_0) = y_0$ — функция

$$y = y_0 e^{p(x-x_0)} \tag{16.32}$$

— стремится к нулю, если $p < 0$, и бесконечно растет по абсолютной величине при $p > 0$. Посмотрим, как ведут себя ошибки простейших численных методов, примененных к модельному уравнению (16.31).

Метод Эйлера (16.11) на модельном уравнении (16.31) (с $f(x, y) = py$ и $f'_y(x, y) = p$) допускает ошибку, которая, согласно (16.19), (16.20), удовлетворяет рекуррентному равенству (иначе, разностному уравнению)

$$\delta_{i+1} = (1 + ph)\delta_i + O(h^2), \tag{16.33}$$

где $i = 0, 1, 2, \ldots$ и $\delta_0 = 0$. Второе слагаемое в (16.33) связано с погрешностью аппроксимации данного дифференциального уравнения (16.31) разностной схемой

$$y_{i+1} = y_i + hpy_i, \tag{16.34}$$

и его влиянием на численную устойчивость процесса (16.34) можно пренебречь (правда, считать его постоянным, т.е. не зависящим от i, можно лишь условно). Характеристическое уравнение (16.27) для (16.33) при $m = 1$ имеет единственный простой корень

$$\lambda = 1 + ph,$$

определяющий фундаментальное решение

$$\Delta_i = (1 + ph)^i$$

*) Впоследствии (см. § 16.7) условие $p \in \mathbf{R}$ будет заменено условием $p \in \mathbf{C}$, и несколько прояснится смысл рассмотрения моделей именно вида (16.31).

соответствующего однородного уравнения. Частным решением уравнения (16.33), согласно форме (16.29), можно считать

$$\bar{\delta}_i = \frac{O(h^2)}{1-(1+ph)} = O(h).$$

Следовательно, представление ошибки, накопленной к $(i+1)$-му шагу реализации (16.34) метода Эйлера (16.11), имеет вид

$$\delta_i = C(1+ph)^i + O(h), \qquad (16.35)$$

где $C = \delta_0 - O(h)$, а под δ_0 может пониматься либо нуль при точном стартовом значении $y_0 = y(x_0)$, либо небольшая ошибка приближенного ввода $y_0 \approx y(x_0)$.

Анализируя поведение ошибки (16.35) при $i \to +\infty$, видим, что ее рост будет ограниченным, если шаг сетки h будет удовлетворять неравенству

$$|1+ph| \le 1. \qquad (16.36)$$

Ясно, что при положительных p это неравенство не может быть выполнено ни при каких $h > 0$; действительно, если решение растет по абсолютной величине (см. (16.32)), то и погрешность получаемого методом Эйлера приближенного решения неизбежно растет. При отрицательных p неравенство (16.36) равносильно условию $0 \le h \le -\frac{2}{p}$, т.е. допустим любой шаг из промежутка $\left(0, -\frac{2}{p}\right]$.

Таким образом, *метод Эйлера* (16.11) *устойчив на модельном уравнении* (16.31), *если в этом уравнении* $p < 0$ *и расчетный шаг метода* $h \le -\frac{2}{p}$. Это ограничение на шаг относит явный метод Эйлера к ***условно устойчивым методам***.

Неявный метод Эйлера (16.12) имеет ошибку, которая в соответствии с (16.22) удовлетворяет разностному уравнению

$$\delta_{i+1} = \frac{\delta_i}{1-ph} + \frac{O(h^2)}{1-ph}.$$

Решение этого уравнения, согласно (16.30) с учетом того, что

$\lambda = \frac{1}{1-ph}$, можно записать так:

$$\delta_i = C\lambda^i + \frac{O(h^2)}{(1-ph)\left(1-\frac{1}{1-ph}\right)} = C\left(\frac{1}{1-ph}\right)^i + O(h).$$

Отсюда видно, что для невозрастания ошибки δ_i с ростом i нужно потребовать выполнения неравенства

$$\frac{1}{|1-ph|} \le 1,$$

что при $ph \ne 1$ равносильно неравенству

$$|1-ph| \ge 1. \qquad (16.37)$$

Для отрицательных p это неравенство выполняется при любых $h > 0$, т.е. *неявный метод Эйлера* ***абсолютно устойчив****). Если же $p > 0$, то в таком случае равносильным (16.37) является неравенство $ph \ge 2$, т.е. рост погрешности будет заведомо ограниченным при условии, что расчетный шаг не слишком мал, а именно, при $h \ge \frac{2}{p}$.

Уточненный метод Эйлера (16.13), приложенный к модельному уравнению (16.31), задается однородным трехточечным разностным уравнением второго порядка

$$y_{i+1} = 2phy_i + y_{i-1}, \qquad (16.38)$$

а ошибка δ_i, накапливаемая этим методом к i-му шагу, согласно (16.23), удовлетворяет неоднородному уравнению

$$\delta_{i+1} = 2ph\delta_i + \delta_{i-1} + O(h^3). \qquad (16.39)$$

Составив характеристическое уравнение вида (16.27) при $m = 2$

$$\lambda^2 = 2ph\lambda + 1$$

и найдя его корни

$$\lambda_1 = ph + \sqrt{1+p^2h^2} \quad \text{и} \quad \lambda_2 = ph - \sqrt{1+p^2h^2},$$

*) Заметим, что понятия абсолютной и условной устойчивости численных процессов решения задач Коши обычно вводят применительно к *асимптотически устойчивым решениям*, т.е. для случая $p < 0$ [158].

видим, что $\lambda_1 > 1$ при $p > 0$, а $\lambda_2 < -1$ при $p < 0$. Значит, соответствующее записи (16.30) общее решение уравнения (16.39)

$$\delta_i = C_1 \lambda_1^i + C_2 \lambda_2^i + O(h^2)$$

с ростом i будет расти по абсолютной величине при любых p и h вне зависимости от начальных условий, определяющих ненулевые постоянные C_1 и C_2.

Таким образом, *уточненный метод Эйлера* (16.13), обладая более высоким порядком аппроксимации, чем явный или неявный методы Эйлера, *является неустойчивым методом*.

Попытаемся понять природу такой неустойчивости метода (16.13).

Сравнивая уравнение (16.38) для каркаса решения и уравнение (16.39) для его ошибки, приходим к выводу, что они имеют одни и те же фундаментальные решения

$$\lambda_1^i = \left(ph + \sqrt{1 + p^2 h^2}\right)^i \quad \text{и} \quad \lambda_2^i = \left(ph - \sqrt{1 + p^2 h^2}\right)^i.$$

Применив к $\left(1 + (ph)^2\right)^{\frac{1}{2}}$ биномиальное разложение и увидев в результате такого разложения несколько первых членов ряда для экспоненты, при малых ph имеем:

$$\lambda_1^i = \left[ph + 1 + \frac{1}{2}(ph)^2 + o\left((ph)^2\right)\right]^i \approx \left(e^{ph}\right)^i = e^{iph}, \qquad (16.40)$$

$$\lambda_2^i = \left[ph - 1 - \frac{1}{2}(ph)^2 + o\left((ph)^2\right)\right]^i \approx \left(-e^{-ph}\right)^i = (-1)^i e^{-iph}. \qquad (16.41)$$

Если теперь рассмотреть точное решение (16.32) модельного уравнения (16.31) на сетке

$$x_0 = 0, \quad x_i = ih,$$

а точнее, фундаментальное решение $Y(x) = e^{px}$ линейного уравнения (16.31), то оказывается, что сеточное фундаментальное решение описывается равенством

$$Y(x_i) = e^{iph}, \qquad (16.42)$$

т.е. совпадает с приближенным представлением одного из фундаментальных решений разностного уравнения (16.38).

Итак, трехточечное разностное уравнение (16.38), являясь уравнением второго порядка, имеет два фундаментальных решения: (16.40) и (16.41), одно из которых является паразитным. При p отрицательных, когда точное решение (16.42) убывает, за счет паразитного фундаментального решения (16.41) происходит рост приближенного решения y_i. Если же p положительно, то

ошибка δ_i, как уже выяснилось, растет, но поскольку в этом случае растет и решение, рост ошибки не страшен (паразитное фундаментальное решение (16.41) при этом затухает, что влечет убывание относительной погрешности приближенного решения).

Возвращаясь к общему случаю уравнения $y' = f(x, y)$, отметим, что в роли параметра p в приведенных и в аналогичных им исследованиях численной устойчивости методов, согласно (16.20), (16.22) и т.п., должны фигурировать значения функции $f'_y(x, y)$. Если можно считать, что $f'_y(x, y) \approx const$, то допустимо использование (кусочно-) постоянных аппроксимаций f'_y и применение фактов теории линейных разностных уравнений с постоянными коэффициентами. В противном случае требуется привлечение более тонких результатов о решениях линейных разностных уравнений с переменными коэффициентами.

16.5. ИССЛЕДОВАНИЕ УСТОЙЧИВОСТИ МНОГОШАГОВЫХ МЕТОДОВ

Будем теперь рассматривать на предмет устойчивости $(2m+1)$-параметрическое семейство линейных m-шаговых методов вида

$$y_{i+1} = \sum_{j=1}^{m} \alpha_j y_{i+1-j} + h \sum_{j=0}^{m} \beta_j f(x_{i+1-j}, y_{i+1-j}) \qquad (16.43)$$

(см. (8.35)) в предположении, что выполняются условия согласованности (8.37), (8.39) его коэффициентов α_i $(i = 1, 2, \ldots, m)$ и β_i $(i = 0, 1, \ldots, m)$, необходимые и достаточные для обеспечения, как минимум, первого порядка точности метода (16.43). Даже на модельном уравнении (16.31) сложно проанализировать накопление погрешностей в численном процессе (16.43) так, как это делалось для простейших разностных схем. Поэтому к исследованию устойчивости многошаговых методов часто применяют упрощенный подход, предложенный в 50-х годах XX века шведским математиком Дальквистом. Суть подхода состоит в том, что нелинейное разностное уравнение (16.43), аппроксимирующее данное дифференциальное уравнение (16.6), в свою очередь, аппроксимируется однородным линейным разностным уравнением с постоянными коэффициентами

$$Y_{i+1} = \sum_{j=1}^{m} \alpha_j Y_{i+1-j}. \qquad (16.44)$$

Оно получается отбрасыванием в уравнении (16.43) второго слагаемого из тех соображений, что сходимость метода (а значит, и поведение каркаса решения, поведение ошибки) изучается при $h \to 0$; наличие множителя h во втором слагаемом позволяет допустить, что оно играет ограниченную роль, если функция f ограничена.

В соответствии с изложенным в двух предыдущих параграфах, поведение решений Y_i разностного уравнения (16.44) и его аппроксимационные свойства по отношению к данному дифференциальному уравнению тесно связаны с величинами корней характеристического уравнения

$$\lambda^m - \alpha_1 \lambda^{m-1} - \ldots - \alpha_{m-1}\lambda - \alpha_m = 0, \qquad (16.45)$$

через которые выражаются фундаментальные решения уравнения (16.44) и аналогичного ему уравнения для ошибок.

Определение 16.4. *Метод* (16.43) *при выполнении условий согласованности* (16.37), (16.39) *называется* ***устойчивым по Дальквисту***, *если все корни характеристического уравнения* (16.45) *по модулю не превосходят единицы и среди корней* λ_k *таких, что* $|\lambda_k| = 1$, *нет кратных**). *Если, кроме того,* m–1 *корней уравнения* (16.45) *по модулю меньше единицы, то метод* (16.43) *называется* ***строго устойчивым*** [138] или ***сильно устойчивым*** [6, 10].

Чтобы осмыслить определение устойчивости по Дальквисту, достаточно вспомнить, что приближенное выражение ошибки δ_i метода (16.43) представляется линейной комбинацией фундаментальных решений δ_{ij} $(j = 1, 2, \ldots, m)$ разностного уравнения

$$\delta_{i+1} = \sum_{j=1}^{m} \alpha_j \delta_{i+1-j},$$

и чтобы не наблюдалось роста ошибки, требуется ограничить единицей корни характеристического уравнения (16.45), через степени которых выражаются фундаментальные решения δ_{ij}. При этом нельзя допустить наличия к р а т н ы х корней с модулями, равными единице, ибо в противном случае неизбежен рост ошибки при $i \to \infty$ из-за наличия в представлении δ_i слагаемых с модулями, пропорциональными $i, i^2, \ldots, i^{k-1}$, где k — показатель кратности корня. Строгая же устойчивость означает, что в

---

*) В [158] это требование называют ***условием корней***, в [3, 188] — ***корневым условием***, а в [12] — ***условием*** $\boldsymbol{\alpha}$.

представлении решения Y_i однородного разностного уравнения (16.44) через фундаментальные решения Y_{ij} $(j = 1, 2, \ldots, n)$ лишь одно слагаемое должно аппроксимировать решение дифференциального уравнения (16.6), а остальные слагаемые должны стремиться к нулю, т.е. паразитные фундаментальные решения должны быть затухающими.

Поскольку данное определение устойчивости не учитывает второе слагаемое в формуле (16.43), с его помощью можно делать лишь грубую отбраковку неустойчивых методов.

Применим определение 16.4 к нескольким изучавшимся ранее методам.

Явный и неявный методы Эйлера (16.11), (16.12) имеют характеристическое уравнение $\lambda - 1 = 0$ с единственным корнем $\lambda_1 = 1$; следовательно они строго устойчивы по Дальквисту (как, кстати, и любой другой одношаговый метод, рассматриваемый в качестве частного случая m-шаговых методов).

Уточненный метод Эйлера (16.13) является двухшаговым методом с характеристическим уравнением $\lambda^2 - 1 = 0$; наличие двух простых корней $\lambda_{1,2} = \pm 1$ говорит об устойчивости этого метода по Дальквисту, но не строгой устойчивости (сравните с более детальным изучением его устойчивости, точнее сказать, неустойчивости, в предыдущем параграфе).

Метод Адамса (явный или неявный)

$$y_{i+1} = y_i + h \sum_{j=0}^{m} \beta_j f\left(x_{i+1-j}, y_{i+1-j}\right)$$

строго устойчив по Дальквисту при любом $m \in \mathbf{N}$, так как его характеристическое уравнение $\lambda^m - \lambda^{m-1} = 0$ имеет один корень, равный единице, а остальные $m - 1$ корней равны нулю.

То же можно сказать и о методе Коуэлла (15.47).

Метод Милна четвертого порядка, как известно, определяется двумя формулами:

прогноза (15.28)

$$\hat{y}_{i+1} = y_{i-3} + \frac{4h}{3}\left(2f_i - f_{i-1} + 2f_{i-2}\right)$$

и коррекции (15.31)

$$y_{i+1} = y_{i-1} + \frac{h}{3}\left(\hat{f}_{i+1} + 4f_i + f_{i-1}\right).$$

Первая из этих формул является четырехшаговой с характеристическим уравнением $\lambda^4 - 1 = 0$, корни которого $\lambda_{1,2} = \pm 1$, $\lambda_{3,4} = \pm i$, а вторая — двухшаговой, для которой соответственно имеем $\lambda^2 - 1 = 0$, $\lambda_{1,2} = \pm 1$. Как видим, та и другая формулы определяют разностный метод, устойчивый по Дальквисту, но не строго устойчивый. Таким образом, метод Милна четвертого порядка, несколько выигрывая у метода прогноза и коррекции (15.22), построенного на основе методов Адамса того же порядка, по точностной характеристике (что отмечалось в конце § 15.3), проигрывает последнему по устойчивости.

Требование устойчивости по Дальквисту должно учитываться при конструировании конкретных методов из семейства (16.43). Это означает, что при подборе параметров в формуле (16.43) следует заботиться не только о том, чтобы она аппроксимировала данное дифференциальное уравнение как можно точнее, но и чтобы соответствующее этой формуле алгебраическое уравнение (16.45) имело ограниченные единицей модули корней и не допускались кратные корни с модулями, равными единице.

Априори $2m+1$ коэффициентов α_j, β_j в методе (16.43) можно подобрать так, чтобы находимые с его помощью приближения y_i аппроксимировали значения решения $y(x_i)$ задачи (16.6) с порядком $2m$, т.е. чтобы этот метод был точен для многочленов степени $2m$ (такие примеры можно найти в § 8.4). Однако доказано, что построенные на этой основе методы наивысшего алгебраического порядка точности являются заведомо неустойчивыми. Этот факт уточняется следующим утверждением [12].

Теорема 16.1. *Пусть s – порядок аппроксимации m-шаговым разностным методом* (16.43) *задачи* (16.6). *Тогда в каждом из следующих случаев*:

а) *метод явный* $(\beta_0 = 0)$ *и* $s > m$;

б) *метод неявный* $(\beta_0 \neq 0)$, *m — нечетное и* $s > m+1$;

в) *метод неявный* $(\beta_0 \neq 0)$, *m — четное и* $s > m+2$

среди корней λ_j характеристического уравнения (16.45) *найдется корень, по модулю больший единицы.*

16.6. ЖЕСТКИЕ УРАВНЕНИЯ И СИСТЕМЫ

Имеются задачи, для которых вопрос об устойчивости или неустойчивости применяемых численных методов стоит наиболее остро и требует большой дифференциации. Речь идет о начальных задачах для дифференциальных уравнений, называемых

жесткими. Такие задачи, возникающие в самых разных прикладных областях, в последние десятилетия являются объектом повышенного внимания специалистов по вычислительной математике и служат тем оселком, на котором оттачиваются понятия, формулировки, методы, алгоритмы, программы.

В литературе можно встретить несколько определений жесткости, отличающихся разным уровнем строгости. При первом знакомстве более важно понять, в чем состоит проблема при численном интегрировании дифференциальных уравнений, выделяющая какие-то из них в разряд жестких, и как ведут себя те или иные методы на таких задачах.

Приведем цитату из посвященной жестким уравнениям монографии К. Деккера и Я. Вервера [60]: «*Сущность явления жесткости состоит в том, что решение, которое нужно вычислить, меняется медленно, однако существуют быстро затухающие возмущения. Наличие таких возмущений затрудняет получение медленно меняющегося решения численным способом*». Попытаемся получить некоторые представления об этом явлении с помощью следующих двух примеров [138].

Пример 16.1. Точным решением задачи Коши

$$y' = -100y + 100, \quad y(0) = 2 \qquad (16.46)$$

является функция

$$y = 1 + e^{-100x}, \qquad (16.47)$$

которая уже при малых положительных x становится близкой к своему предельному значению $y = 1$. Это означает, что для получения численного решения на достаточно большом промежутке $[0, b]$ естественно с мелким шагом построить решение в его переходной фазе (в правой окрестности нуля, т.е. там, где оно быстро меняется), а затем продолжить численный процесс с крупным шагом в области малых изменений решения.

Посмотрим, на что можно рассчитывать, применяя к задаче (16.46) (отличающейся от модельной задачи (16.31) лишь наличием свободного члена) простейшие разностные схемы, устойчивость которых изучалась в § 16.4.

Расчетная формула условно устойчивого метода Эйлера (16.11) для (16.46) имеет вид

$$y_{i+1} = (1 - 100h)y_i + 100h; \quad i = 0, 1, \ldots; \quad y_0 = 2. \qquad (16.48)$$

В соответствии с условием устойчивости (16.36) расхождение между y_i и $y(x_i)$ не будет расти, если расчетный шаг удовлетворяет неравенству $|1 - 100h| \le 1$, т.е. если $h \le 0.02$. Взяв $h = 0.02$, из (16.48) получаем численный процесс

$$y_{i+1} = 2 - y_i; \quad i = 0, 1, \ldots; \quad y_0 = 2,$$

порождающий последовательность

$$y(0)=y_0=2,\ y(0.02)\approx y_1=0,\ y(0.04)\approx y_2=2,\ y(0.06)\approx y_3=0,\ \dots .$$

Как видим, решение разностного уравнения получается ограниченным, но нельзя считать, что оно удовлетворительно приближает решение (16.47) задачи (16.46) (имеет место *«четно-нечетная болтанка»* [9]). При $h=0.01$ (это середина допустимого для h интервала устойчивости $(0, 0.02)$ процесс (16.48) приобретает вид

$$y_0=2,\quad y_{i+1}=1\quad \forall i\in N_0,$$

показывающий, что первая же вычисляемая точка y_1 ($\approx y(0.01)$) попадает на асимптоту решения (16.47), и последующие вычисления не изменяют значений приближенного решения. Несмотря на устойчивость этого численного процесса и стремящуюся к нулю ошибку при $i\to\infty$, он совсем не описывает переходный процесс. Существенно более мелкий шаг, например, $h=0.001$ вполне удовлетворительно характеризует начальное изменение решения, что видно из получающейся при этом последовательности

$$y_0=2,\quad y(0.001)\approx y_1=1.9,\quad y(0.002)\approx y_2=1.81,\ \dots,$$

но потребует больших вычислительных затрат (и, как следствие, повлечет накопление ошибок округлений) при построении приближенного решения на промежутке, достаточно большом по сравнению с величиной шага h.

Применение здесь абсолютно устойчивого неявного метода Эйлера (16.12) означает проведение вычислений по формуле

$$y_{i+1}=\frac{100h+y_i}{1+100h};\quad i=0,1,\dots;\quad y_0=2. \tag{16.49}$$

Полагая здесь для удобства счета $h=0.09$, получаем следующую последовательность приближенных ординат решения (16.47):

$$y_0=2,\ y(0.09)\approx y_1=1.1,\ y(0.18)\approx y_2=1.01,\ y(0.27)\approx y_3=1.001,\ \dots.$$

Даже при очень крупном шаге $h=0.99$ определяемая посредством (16.49) последовательность

$$y_0=2,\quad y_1=1.01,\quad y_2=1.0001,\ \dots$$

качественно верно определяет поведение решения (16.47) задачи (16.46). Так что для получения приближенного решения данной задачи неявным методом Эйлера с нужной точностью при выборе расчетного шага h достаточно позаботиться лишь об адекватной точности аппроксимации дифференциального уравнения разностным.

Неустойчивая разностная схема (16.13), иначе, уточненный метод Эйлера, в данном примере приобретает вид

$$y_{i+1}=y_{i-1}-200hy_i+200h;\quad i=1,2,\dots, \tag{16.50}$$

где, кроме известной начальной точки $y_0=y(0)=2$, требуется задание еще одной точки y_1. Положим $h=0.01$ и, соответственно, за y_1 примем значение

$$y_1=1.37\quad \left(\approx y(0.01)=1+\frac{1}{e}\right).$$

В таком случае вычисления по рекуррентной формуле (16.50) продолжают эту последовательность следующим образом:

$$y_2=1.26,\ y_3=0.85,\ y_4=1.56,\ y_5=-0.27,\ y_6=4.10,\ \dots.$$

Не вызывает сомнений ее непригодность для приближенного описания поведения решения (16.47). Взяв в 100 раз меньший шаг $h=0.0001$, при котором погрешность аппроксимации дифференциальной задачи (16.46) дискретной задачей (16.50), согласно (16.9), есть $O(10^{-8})$, для тех же начальных элементов последовательности $y_0=2$ и $y_1=1.37$ имеем продолжение

$$y_2=1.993,\ y_3=1.350,\ y_4=1.986,\ y_5=1.330,$$

$$y_6=1.979,\ y_7=1.311,\ y_8=1.973,\ \dots.$$

Пульсирующий характер последовательностей с достаточно большими амплитудами пульсаций может быть объяснен поведением паразитной составляющей решения разностного уравнения (16.50), о чем говорилось при анализе причин неустойчивости уточненного метода Эйлера в § 16.4.

Пример 16.2. Уравнение второго порядка

$$y''+101y'+100y=0 \tag{16.51}$$

с начальными условиями

$$y(0)=1.01,\quad y'(0)=-2$$

имеет решением функцию

$$y(x)=0.01e^{-100x}+e^{-x}. \tag{16.52}$$

Первое из слагаемых, составляющих эту функцию, является быстрозатухающим, а второе изменяется сравнительно плавно; проследить за поведением этих двух слагаемых на промежутке $[0, 0.1]$ можно по следующей таблице их приближенных значений:

x	$Y_1=0.01e^{-100x}$	$Y_2=e^{-x}$
0	0.01	1
0.00001	0.009999	0.99999
0.0001	0.0099	0.9999
0.001	0.009	0.999
0.01	0.004	0.99
0.1	0.0000004	0.9

Очевидно, за пределами промежутка $[0, 0.1]$, характеризующего переходную фазу, вклад первого слагаемого Y_1 в решение (16.52) задачи (16.51) ничтожен. Это означает, что при последовательном получении приближенных численных значений решения $y(x)$ промежуток $[0, 0.1]$ нужно проходить с маленьким шагом, а затем, чтобы уменьшить вычислительные затраты и погрешность округлений, расчетный шаг должен быть увеличен. Всякие ли вычислительные методы позволяют произвести укрупнение шага после построения приближенного решения в переходной фазе? Ответ на этот вопрос будет отрицательным.

Действительно, как отмечалось в § 15.6, для построения численного решения задачи (16.51) формально можно применять разные численные методы решения задачи Коши для уравнений первого порядка в векторной форме, сведя ее к системе

$$\begin{cases} y' = z, & y(0) = 1.01, \\ z' = -100y - 101z, & z(0) = -2. \end{cases} \quad (16.53)$$

Использование здесь условно устойчивых явных методов, например, таких как метод Эйлера, накладывает ограничение на величину шага, обусловленное как раз необходимостью устойчивой аппроксимации быстрозатухающей составляющей решения.

Вообще, переход к системе дифференциальных уравнений позволяет взглянуть на проблему жесткости с несколько иных позиций.

Пусть n-мерная система

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} \quad (16.54)$$

имеет асимптотически устойчивое решение

$$\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))^T.$$

Если его компоненты, т.е. функции $x_1(t), \ldots, x_n(t)$, существенно различаются по скорости своего изменения на промежутке $[t_0, T]$, на котором решается задача Коши для (16.54), то применение здесь условно устойчивых численных процессов требует интегрирования с таким малым шагом, какой обеспечивает устойчивое вычисление самой быстрозатухающей компоненты (ведь шаг h — величина скалярная, общая для всех компонент). Следовательно, жесткость системы дифференциальных уравнений зависит от того, насколько сильно разнится поведение компонент вектора-решения при условии его асимптотической устойчивости. В свою очередь, в случае постоянной матрицы $\mathbf{A}$ в системе (16.54) это различие в скорости изменения функций $x_1(t), \ldots, x_n(t)$ связано с тем, насколько сильно различаются собственные числа $\lambda_1, \ldots, \lambda_n$ матрицы $\mathbf{A}$. Обоснование такой связи можно найти, например, в [6, 158]; возвращаясь же к примеру 16.2 и найдя собственные числа $\lambda_1 = -100$, $\lambda_2 = -1$ матрицы $\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -100 & -101 \end{pmatrix}$ системы (16.53), убеждаемся, по крайней мере, в рациональности такого суждения.

Отсюда приходим к следующему определению жесткой системы [158].

Определение 16.5. *Система* (16.54) *с постоянной $n \times n$-матрицей* $\mathbf{A}$ *называется* ***жесткой***, *если собственные числа* λ_k $(k = 1, 2, \ldots, n)$ *матрицы* $\mathbf{A}$ *удовлетворяют следующим условиям*:

1) $\operatorname{Re}\lambda_k < 0 \quad \forall k \in \{1, 2, \ldots, n\}$;

2) ***Число жесткости*** $g := \dfrac{\max\{|\operatorname{Re}\lambda_k|\}}{\min\{|\operatorname{Re}\lambda_k|\}}$ *велико.*

Нестрогость, заключенная в последнем слове данного определения, сродни той, которая присутствует при введении понятия «плохая обусловленность матрицы» (см. гл. 1); избавиться от нее можно, лишь рассматривая конкретную задачу.

В случае линейной системы с переменными коэффициентами, т.е. при $\mathbf{A} = \mathbf{A}(t)$ в (16.54), собственные числа λ_k и, соответственно, число жесткости g являются функциями от t, и определение 16.5 жесткой системы может быть переформулировано следующим образом.

Определение 16.6. *Система*

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}, \quad t > 0$$

называется ***жесткой на интервале*** $(0, T)$, *если при всех* $t \in (0, T)$

$$\operatorname{Re}\lambda_k(t) < 0 \quad \forall k \in \{1, 2, \ldots, n\}$$

и число $\sup\limits_{t \in (0,T)} g(t)$ *велико.*

В основе определения жесткости нелинейной системы

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}, t)$$

лежит определение (16.6), роль матрицы $\mathbf{A}(t)$ в котором отводится матрице частных производных $\dfrac{\partial f_i}{\partial x_j}$ $(i, j = 1, 2, \ldots, n)$, где f_i и x_i — компоненты вектор-функций $\mathbf{F}$ и $\mathbf{x}$ соответственно.

Заметим, что требование асимптотической устойчивости точного решения системы дифференциальных уравнений, обеспечивающееся выполнением первого условия приведенных определений жесткости, иногда заменяют более слабым условием асимптотической устойчивости не всех, а только быстрозатухающих компонент вектора-решения, что расширяет класс жестких систем. С другой стороны, промежуток, на котором действуют быстрозатухающие возмущения, часто не относят к промежутку жесткости [60] (например, в примере 16.1 таковым является промежуток $[0, 0.1]$).

16.7. A- И $A(\alpha)$-УСТОЙЧИВОСТЬ. ЧИСТО НЕЯВНЫЕ МЕТОДЫ

Как следует из материала предыдущего параграфа, жесткие уравнения предъявляют жесткие требования к устойчивости численных методов, применяемых для их решения. А именно, при получении асимптотически устойчивого решения жесткой задачи Коши ошибка разностного метода не должна расти при любом шаге, т.е. метод должен быть безусловно устойчивым. Чтобы оформить сказанное более четко, дадим сначала определение области устойчивости [158].

Определение 16.7. ***Областью устойчивости*** *разностного метода* (16.43) *решения начальной задачи* (16.6) *называется множество всех точек комплексной плоскости, определяемой комплексной переменной* $\mu = ph$, *для которых этот метод, примененный к модельному уравнению* (16.31), *устойчив, т.е. обеспечивает невозрастание ошибки.*

Теперь сформулируем определение устойчивости метода, ориентированное на применение к жестким задачам.

Определение 16.8. *Разностный метод* (16.43) *называется* ***А-устойчивым****, если его область устойчивости содержит левую полуплоскость комплексной плоскости, определяемой переменной* $\mu = ph$.

Обращаясь к простейшим разностных схемам, сразу отметим, что об A-устойчивости явного метода Эйлера (16.11) не может быть и речи. Действительно, согласно проведенным в § 16.4 исследованиям, для невозрастания ошибки этого метода требуется выполнение неравенства (см. (16.36))

$$|\mu + 1| \le 1,$$

что реализуется в круге радиуса 1 с центром в точке $(-1; 0)$ комплексной плоскости с осями Ou и Ov (рис. 16.2а), где

$$u = \operatorname{Re}\mu = \operatorname{Re}(ph), \quad v = \operatorname{Im}\mu = \operatorname{Im}(ph).$$

Этим неравенством задается его область устойчивости, составляющая лишь малую часть левой полуплоскости плоскости Ouv.

Устойчивость неявного метода Эйлера (16.12), в соответствии с неравенством (16.37), связана с выполнением неравенства

$$|\mu - 1| \ge 1,$$

которое определяет на комплексной плоскости Ouv внешнюю

часть такого же круга, но круг этот расположен в правой полуплоскости (рис. 16.2б). Как видим, область устойчивости неявного метода Эйлера целиком содержит левую полуплоскость; следовательно, *неявный метод Эйлера А-устойчив.*

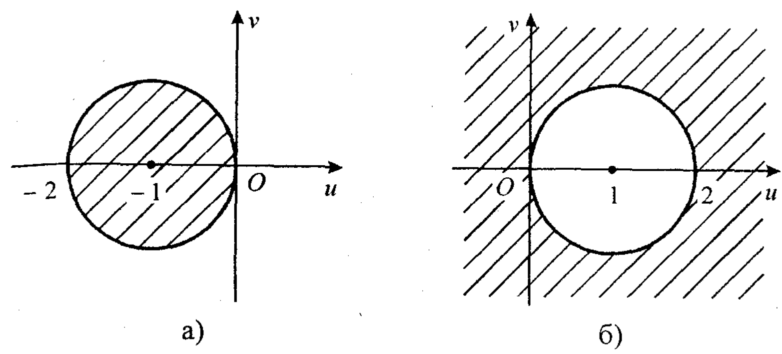


Рис. 16.2. Области устойчивости явного (а) *и неявного* (б) *методов Эйлера*

Рассмотрим на предмет A-устойчивости метод трапеций (14.15) (он же — метод Адамса-Моултона второго порядка (15.19)). На модельном уравнении (16.31) его расчетная формула имеет вид

$$y_{i+1} = y_i + \frac{ph}{2}[y_i + y_{i+1}], \quad i = 0, 1, \ldots. \tag{16.55}$$

Те или другие источники получения разностной схемы (16.55) позволяют считать, что подстановка в нее точного значения $y(x_i)$ вместо y_i сопровождается ошибкой $O(h^3)$, т.е. имеет место равенство

$$y(x_{i+1}) = y(x_i) + \frac{ph}{2}\left[y(x_i) + y(x_{i+1})\right] + O(h^3).$$

Вычитая равенство (16.55) из последнего равенства, приходим к следующему двухточечному разностному уравнению первого порядка относительно ошибки $\delta_i := y(x_i) - y_i$:

$$\left(1 - \frac{ph}{2}\right)\delta_{i+1} = \left(1 + \frac{ph}{2}\right)\delta_i + O(h^3).$$

Отсюда видим, что невозрастание ошибки δ_i гарантируется при выполнении неравенства

$$\left|\frac{2 + ph}{2 - ph}\right| \le 1.$$

Полагая здесь $\mu = ph = u + iv$, $\mu \neq 2$, имеем

$$|2+u+iv| \le |2-u-iv| \quad \Leftrightarrow \quad \sqrt{(2+u)^2+v^2} \le \sqrt{(2-u)^2+v^2}$$

$$\Leftrightarrow \quad u \le 0 \quad \Leftrightarrow \quad \operatorname{Re}\, \mu \le 0.$$

Таким образом, область устойчивости метода трапеций точно совпадает с левой полуплоскостью комплексной плоскости чисел $\mu = ph$; следовательно, *метод трапеций A-устойчив.*

Докажем также ***A*-устойчивость неявного двухшагового разностного метода второго порядка (16.15)**, выведенного в § 16.2.

Аналогично предыдущему, легко получаем разностное уравнение для ошибки каркаса решения модельного уравнения (16.31). Оно имеет вид

$$(3-2ph)\delta_{i+1} - 4\delta_i + \delta_{i-1} + O(h^3) = 0.$$

Его характеристическое уравнение

$$(3-2\mu)\lambda^2 - 4\lambda + 1 = 0 \tag{16.56}$$

имеет корнями числа

$$\lambda_{1,2} = \frac{2 \pm \sqrt{1+2\mu}}{3-2\mu},$$

непосредственное оценивание которых, точнее, построение области, содержащей только те точки, при которых $|\lambda_{1,2}| \le 1$, вызывает определенные затруднения. Изберем другой путь [158].

Попытаемся выяснить вид г р а н и ц ы области устойчивости, рассматривая равенство (16.56) при таких комплексных λ, модуль которых равен единице. При этом будем пользоваться как тригонометрической, так и показательной формами комплексного числа, т.е. тем, что при $|\lambda| = 1$ по формуле Эйлера

$$\lambda = \cos\varphi + i\sin\varphi = e^{i\varphi}. \tag{16.57}$$

Выражая из (16.56) текущую комплексную переменную μ, с помощью равенства (16.57) получаем:

$$\mu = \frac{1}{2\lambda^2} - \frac{2}{\lambda} + \frac{3}{2} = \frac{1}{2}e^{-2i\varphi} - 2e^{-i\varphi} + \frac{3}{2} =$$

$$= \frac{1}{2}\cos 2\varphi - \frac{1}{2}i\sin 2\varphi - 2\cos\varphi + 2i\sin\varphi + \frac{3}{2} = u + iv,$$

где

$$u := \frac{3}{2} - 2\cos\varphi + \frac{1}{2}\cos 2\varphi = (1-\cos\varphi)^2 \ge 0, \tag{16.58}$$

$$v := 2\sin\varphi - \frac{1}{2}\sin 2\varphi = \sin\varphi(2-\cos\varphi). \tag{16.59}$$

Найденные выражения u и v текущей точки μ границы области устойчивости можно рассматривать как ее параметрические уравнения, в которых параметром служит полярный угол φ. Из (16.58) сразу видно, что граница находится в правой полуплоскости, а исследуя определенную в (16.59) функцию v на экстремум, выясняем, что существуют два значения параметра $\varphi_{1,2} = \arccos\frac{1 \pm \sqrt{3}}{2}$, которым соответствуют экстремальные значения $v \approx \pm 2$ при $u \approx 1.85$. Найдя еще дополнительно две характерные точки границы $(0;0)$ и $(4;0)$, схематично изображаем замкнутую линию, определяемую параметрическими уравнениями (16.58), (16.59) (рис. 16.3). Эта линия делит всю комплексную плоскость на две области: в одной из них модули корней уравнения (16.56) меньше единицы, в другой — хотя бы один из корней больше единицы. Подставив в (16.56), например, $\mu = -\frac{1}{2}$, получаем уравнение $4\lambda^2 - 4\lambda + 1 = 0$ с корнями $\lambda_{1,2} = \frac{1}{2} < 1$. Следовательно, область устойчивости метода (16.15) лежит за пределами построенной замкнутой линии и содержит левую полуплоскость.

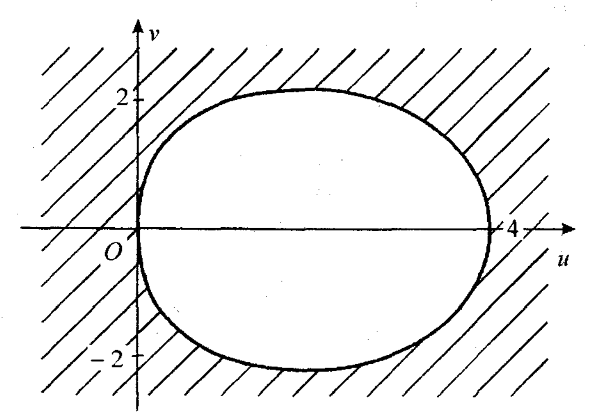


Рис. 16.3. Область устойчивости разностной схемы
$$y_{i+1} = \frac{4}{3}y_i - \frac{1}{3}y_{i-1} + \frac{2h}{3}f(x_{i+1}, y_{i+1})$$

К сожалению, класс A-устойчивых методов весьма узок. Доказано, что *среди линейных многошаговых методов* (16.43) *нет явных A-устойчивых методов, а порядок неявных A-устойчивых методов этого семейства* (16.43) *не может быть больше двух.*

В связи с этим представляют интерес формулировки определений более слабой устойчивости, которые расширяли бы множество разностных методов, хотя бы ограниченно пригодных для численного решения жестких задач.

Прямым обобщением A-устойчивости является $A(\alpha)$-устойчивость.

Определение 16.9. *Разностный метод* (16.43) *называется* ***$A(\alpha)$-устойчивым****, если существует угол* $\alpha \in \left(0, \frac{\pi}{2}\right]$ *такой, что область устойчивости метода* (рис. 16.4) *содержит сектор комплексной плоскости переменных $\mu = ph$, определяемый неравенством*

$$|\arg(-\mu)| < \alpha$$

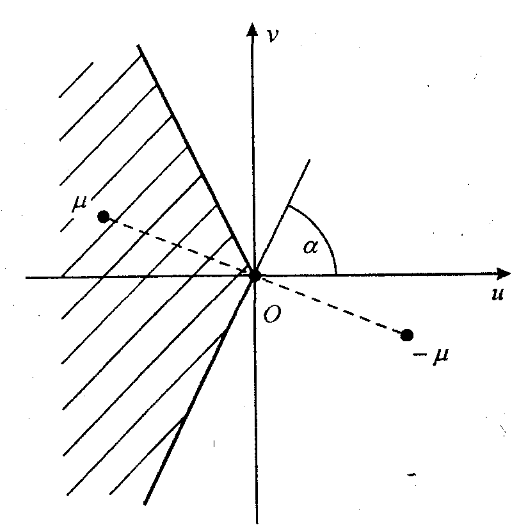


Рис. 16.4. *Изображение области (заштрихована), которая должна содержаться в области устойчивости $A(\alpha)$-устойчивого метода*

Очевидно, что A-устойчивость — это $A\left(\frac{\pi}{2}\right)$-устойчивость.

$A(\alpha)$-устойчивые методы можно найти в подсемействе семейства m-шаговых разностных методов (16.43), определяемом формулой

$$\sum_{j=0}^{m} \alpha_j y_{i+1-j} = hf(x_{i+1}, y_{i+1}). \tag{16.60}$$

Счет по формулам вида (16.60) не использует значений производных решения в предыдущих узлах, и поэтому такие методы называют ***чисто неявными методами***. Источником получения конкретных $A(\alpha)$-устойчивых чисто неявных методов вида (16.60) служит применение к аппроксимации производной в равенстве

$$y'(x_{i+1}) = f(x_{i+1}, y_{i+1}) \tag{16.61}$$

формул дифференцирования назад*), т.е. несимметричных формул, аппроксимирующих $y'(x_{i+1})$ с разным порядком точности с помощью значений y_{i+1}, y_i, y_{i-1} и т.д.

Так, аппроксимация $y'(x_{i+1})$ в (16.61) разностными отношениями

$$\frac{1}{h}(y_{i+1} - y_i) \quad \text{и} \quad \frac{1}{2h}(3y_{i+1} - 4y_i + y_{i-1})$$

приводит к изученным нами $A\left(\frac{\pi}{2}\right)$-устойчивым чисто неявным методам первого (16.12) и второго (16.15) порядков. Аналогично, аппроксимация $y'(x_{i+1})$ по формулам, например, третьего и четвертого порядков приводит к чисто неявным методам

$$11y_{i+1} - 18y_i + 9y_{i-1} - 2y_{i-2} = 6hf(x_{i+1}, y_{i+1}) \tag{16.62}$$

и

$$25y_{i+1} - 48y_i + 36y_{i-1} - 16y_{i-2} + 3y_{i-3} = 12hf(x_{i+1}, y_{i+1}), \tag{16.63}$$

которые, как доказано, являются $A(\alpha)$-устойчивыми с $\alpha \approx 1.544$ рад $(\approx 88^\circ)$ для метода третьего порядка (16.62) и с $\alpha = arctg\sqrt{6} \approx 1.278$ рад $(\approx 68^\circ)$ для метода четвертого порядка (16.63) [158]. С повышением порядка метода дифференцирова-

*) В связи с чем и выводимые таким образом методы численного решения задач Коши часто называют ***методами дифференцирования назад*** [188].

ния назад значения угла α уменьшаются, и при порядке выше шестого $A(\alpha)$-устойчивых методов среди семейства многошаговых методов (16.60) не обнаруживается ни для каких $\alpha > 0$ [185].

Заметим, что при применении чисто неявных методов к численному интегрированию жестких задач соответствующие используемому методу разностные уравнения, в общем случае нелинейные относительно очередного значения y_{i+1}, должны решаться достаточно точно. Здесь рекомендуется привлекать быстросходящиеся итерационные методы типа метода Ньютона (см. гл. 5–7, а также [6, 188]).

Вернемся к началу этого параграфа и посмотрим, что означают A- и $A(\alpha)$-устойчивость разностных методов, применяемых именно к жестким **системам** дифференциальных уравнений.

Если некоторый метод A-устойчив, то он пригоден для формального нахождения приближенного решения модельного уравнения $y' = py$ с любым шагом $h > 0$ при любых комплексных p с отрицательными вещественными частями, а это в свете наблюдений предыдущего параграфа означает, что шаг h может быть подобран так, чтобы система (16.54) решалась устойчиво при любом разбросе собственных значений ее матрицы $\mathbf{A}$, т.е. практически при любой величине числа жесткости g. Если же метод $A(\alpha)$-устойчив с $\alpha < \frac{\pi}{2}$, то допустимость этого метода для решения системы (16.54) можно связать с тем, вписываются ли собственные числа λ_k ее матрицы, а точнее, величины $\lambda_k h$, в определяемый данным α сектор комплексной плоскости (см. определение 16.9).

УПРАЖНЕНИЯ

16.1. Оцените сверху величину $|y(1) - y_n|$, где $y(1)$ — точное решение задачи

$$y' = \sin(xy), \quad y(0) = 0$$

в точке $x = 1$, а y_n — приближение к нему, полученное методом Эйлера с шагом $h = \frac{1}{n}$.

16.2. Продолжите серию разностных схем (16.11)–(16.15), используя формулы численного дифференцирования третьего порядка точности.

16.3. Исследуйте на устойчивость по Дальквисту:

а) явный и неявный двухшаговые разностные методы второго порядка (16.14) и (16.15);

б) явный двухшаговый метод третьего порядка (15.46);

в) явный и неявный методы, выведенные при выполнении упр.15.3.

16.4. Исследуйте на модельном уравнении $y' = py$ поведение ошибки метода Адамса–Башфорта второго порядка (15.12). Что можно сказать о его устойчивости?

16.5. Проанализируйте выражения корней характеристического уравнения для однопараметрического семейства двухшаговых методов третьего порядка (15.45), выделите из этого семейства наиболее устойчивый по Дальквисту метод. Встречался ли этот метод ранее? Как он называется?

16.6. Убедитесь, что $A(\alpha)$-устойчивый чисто неявный метод третьего порядка (16.62) является строго устойчивым по Дальквисту.

16.7. При каком ограничении на шаг можно рассчитывать на устойчивое решение задачи

$$\begin{cases} \dot{x} = -11x + y, & x(0) = 0, \\ \dot{y} = -7x - 3y, & y(0) = 0 \end{cases}$$

методом Эйлера?

16.8. Не находя области устойчивости чисто неявного метода второго порядка (16.15), покажите, что для всех точек комплексной плоскости чисел $\mu = ph$ выполняется условие $|\lambda_{1,2}| \le 1$, где $\lambda_{1,2}$ — корни характеристического уравнения (16.56).

ГЛАВА 17 МЕТОДЫ ПРИБЛИЖЕННОГО РЕШЕНИЯ КРАЕВЫХ ЗАДАЧ ДЛЯ ОБЫКНОВЕННЫХ ДИФФЕРЕНЦИАЛЬНЫХ УРАВНЕНИЙ

Рассматривается несколько способов приближенного решения двухточечных линейных краевых задач для дифференциальных уравнений второго порядка.

Сначала изучаются методы, позволяющие применить накопленный арсенал методов численного решения начальных задач. Далее на основе простейших аппроксимаций производных в узлах сетки строится удобный для автоматизированных вычислений метод конечных разностей получения каркасов решений и изучается его численная устойчивость. Из приближенно-аналитических методов здесь представлены метод коллокации и метод Галёркина. Показывается, как можно подбирать требуемые этими методами системы базисных функций в случае произвольных линейных краевых условий. В развитие метода Галёркина выводятся расчетные формулы метода конечных элементов, где в качестве базисных функций используются B-сплайны. Подобно конечноразностному методу его применение сводится к решению линейных алгебраических систем с трехдиагональными матрицами коэффициентов.

17.1. ПОСТАНОВКА ЗАДАЧИ. КЛАССИФИКАЦИЯ ПРИБЛИЖЕННЫХ МЕТОДОВ

Будем рассматривать двухточечные краевые задачи для обыкновенных дифференциальных уравнений в т о р о г о порядка. Общий вид таких задач

$$\begin{aligned} &F(x, y, y', y'') = 0, \quad x \in [a, b], \\ &\varphi_1(y(a), y'(a)) = 0, \quad \varphi_2(y(b), y'(b)) = 0, \end{aligned} \qquad (17.1)$$

где F, φ_1, φ_2 — заданные функции определенной гладкости. Наиболее употребительны и лучше всего изучены л и н е й н ы е краевые задачи, т.е. задачи вида (17.1), в которых F, φ_1 и φ_2 — линейные функции. Для определенности будем считать основным объектом дальнейшего изучения ***линейную краевую задачу***

$$L[y] := y'' + p(x)y' + q(x)y = f(x), \quad x \in [a, b], \qquad (17.2)$$

$$l_a[y] := \alpha_0 y(a) + \alpha_1 y'(a) = A, \qquad (17.3)$$

$$l_b[y] := \beta_0 y(b) + \beta_1 y'(b) = B, \qquad (17.4)$$

где к коэффициентам ***краевых условий*** (17.3), (17.4) предъявляется требование

$$|\alpha_0| + |\alpha_1| \neq 0, \quad |\beta_0| + |\beta_1| \neq 0, \qquad (17.5)$$

а функции $p = p(x)$, $q = q(x)$ и $f = f(x)$ в уравнении (17.2) должны быть такими, чтобы данная задача имела единственное решение $y = y(x)$ в заданном функциональном пространстве. Краевые условия (17.3) , (17.4) определяют так называемую ***третью*** или, иначе, ***смешанную краевую задачу*** для уравнения (17.2), содержащую в себе ***первую*** (когда $\alpha_1 = \beta_1 = 0$) или ***вторую*** (при $\alpha_0 = \beta_0 = 0$) краевые задачи. Оставив в стороне важный случай ***периодической краевой задачи***, когда вместо (17.3), (17.4) выставляются условия $y(a) = y(b)$, $y'(a) = y'(b)$, будем конструировать методы приближенного решения с м е ш а н н о й задачи (17.2)–(17.4), лишь изредка выделяя случаи первой и (или) второй задач, если это окажется существенным.

Как известно, точное (аналитическое) решение краевых задач вызывает еще бóльшие трудности, чем решение задач Коши. Отсюда — повышенный интерес и большое разнообразие приближенных методов решения таких задач. По типу представления результатов приближенного решения методы можно разделить на две группы: приближенно-аналитические, дающие приближенное решение краевой задачи на отрезке $[a, b]$ в виде некоторой конкретной функции, и собственно численные или сеточные методы, дающие каркас приближенного решения на заданной на $[a, b]$ сетке*). По идейной основе приближенных методов их можно классифицировать следующим образом:

1) методы сведения к задаче Коши (метод пристрелки, метод дифференциальной прогонки, метод редукции);

2) метод конечных разностей;

3) метод балансов или интегро-интерполяционный метод;

4) метод коллокации;

5) проекционные методы (моментов, Галёркина);

6) вариационные методы (наименьших квадратов, Ритца);

7) проекционно-разностные методы (метод конечных элементов);

8) методы сведения к интегральным уравнениям Фредгольма и др.

---

*) При этом некоторые методы первой группы также предполагают использование сетки.

Методы 4...6 из вышеперечисленных приводят к приближенному решению в виде функции заданного семейства (линейной комбинации некоторой системы линейно независимых функций), методы 1...3 и 7 генерируют таблицы численных значений приближенного решения, в методах 8 возможны варианты. Как правило, чисто сеточные методы являются более простыми и позволяют технически легко строить каркас решения на заданной сетке с наперед заданной точностью, контролируемой, например, по принципу Рунге. Однако и приближенно-аналитические методы имеют свои достоинства, одно из которых очевидно — это лаконичность функционального представления решения; другое же их достоинство состоит в том, что некоторые методы этой группы позволяют получать хорошие приближения к обобщенным решениям краевой задачи, когда она не имеет единственного решения в классическом смысле.

Ниже будут рассмотрены идеи и некоторые реализации методов 1, 2, 4, 5, 7 приведенного списка; при желании познакомиться с другими методами или с иным освещением перечисленных читатель может обратиться к другим учебным пособиям, например, к [3, 13, 20, 62, 100, 131, 158, 181].

17.2. МЕТОДЫ СВЕДЕНИЯ КРАЕВЫХ ЗАДАЧ К НАЧАЛЬНЫМ

Имея в виду, что нами изучено много способов приближенного решения задач Коши для ОДУ, можно считать краевую задачу п р и н ц и п и а л ь н о решенной, если ее удастся преобразовать к эквивалентной начальной задаче*). Известно несколько приемов, с помощью которых можно это сделать. Эти приемы приводят к необходимости решать не одну, а несколько задач Коши разной сложности при различных ограничениях на параметры исходной задачи. Рассмотрим вкратце три таких приема.

---

*) Автор считает уместным напомнить здесь весьма популярную в математических кругах притчу о различии подходов к решению прикладных задач двух групп людей, представителей которых будем условно называть «практик» и «теоретик».

Практику и теоретику дали по доске с забитым п о ш л я п к у гвоздем и инструменты: молоток и клещи, предложив вытащить гвоздь. Повозившись, оба справились с задачей (не будем уточнять, кто быстрее). Второе задание состояло в том, чтобы вытащить гвоздь, вбитый в доску лишь н а п о л о в и н у. Практик взял клещи и вытащил гвоздь. Теоретик взял молоток и со словами: «Задача сведена к предыдущей», забил гвоздь по шляпку.

Метод пристрелки. Пусть требуется найти приближенное решение уравнения (в общем случае нелинейного)

$$y'' = f(x, y, y'), \quad x \in [a, b] \qquad (17.6)$$

с краевыми условиями первого рода

$$y(a) = A, \quad y(b) = B. \qquad (17.7)$$

Зададимся некоторым числом C_1 и будем рассматривать задачу Коши для уравнения (17.6) с н а ч а л ь н ы м и условиями

$$y(a) = A, \quad y'(a) = C_1.$$

Предположим, что на отрезке $[a, b]$ построено приближенное решение $y = y_1(x, C_1)$ этой начальной задачи. Сравнение значения $y_1(b, C_1)$ полученного решения в точке $x = b$ с заданным, согласно (17.7), значением B дает информацию для корректирования угла наклона касательной к решению $y = y_2(x, C_2)$ новой начальной задачи с условиями

$$y(a) = A, \quad y'(a) = C_2$$

так, чтобы уменьшить разницу $B - y_2(b, C_2)$. На этой основе могут выстраиваться те или иные стратегии варьирования задаваемых значений $y'(a) = C_1, C_2, C_3, \ldots$, обеспечивающих, по возможности, наиболее быструю практическую сходимость к числу B последовательности приближений $y_k(b, C_k)$ $(k = 1, 2, 3, \ldots)$*). Если при заданном $\varepsilon > 0$ и некотором $k = n$ будет выполнено неравенство $|B - y_n(b, C_n)| \le \varepsilon$, за искомое приближенное решение краевой задачи (17.6)–(17.7) принимается функция $y = y_n(x, C_n)$ (дискретная или непрерывная).

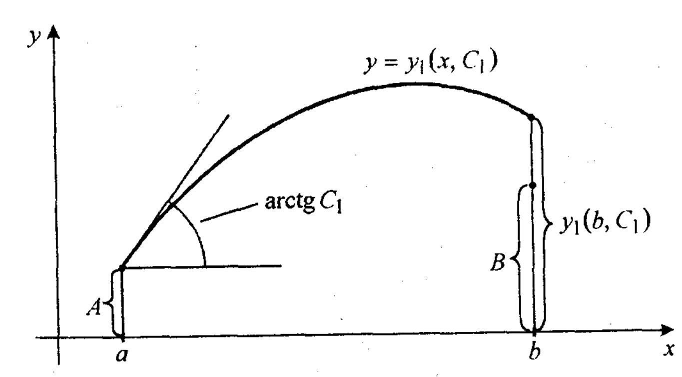


Рис. 17.1. Геометрическая интерпретация одного шага метода пристрелки

---

*) См., например, [3, 178].

Схематично описанный способ получения приближенного решения первой краевой задачи путем последовательного решения нескольких задач Коши для того же уравнения называют ***методом пристрелки*** (или ***стрельбы***). Такое название становится естественным, если рассмотреть геометрическую (или, если угодно, баллистическую) интерпретацию метода (рис. 17.1).

Метод редукции. Будем искать решение $y = y(x)$ линейной краевой задачи (17.2)–(17.4) в виде

$$y = Cu(x) + v(x), \tag{17.8}$$

где C и $u = u(x)$, $v = v(x)$ — некоторые постоянная и функции, условия на которые будут накладываться ниже.

Сначала потребуем, чтобы функция (17.8) была решением уравнения (17.2) при л ю б о м значении постоянной C, т.е. чтобы имело место

$$L[Cu + v] = f \quad \forall C \in \mathrm{R}.$$

В силу линейности заданного в (17.2) дифференциального оператора L, последнее можно переписать в виде

$$CL[u] + L[v] = f \qquad \forall C \in \mathrm{R}, \tag{17.9}$$

а так как постоянная C должна быть произвольной, то уравнение (17.9) равносильно системе

$$\begin{cases} L[u] = 0, \\ L[v] = f. \end{cases} \tag{17.10}$$

Теперь воспользуемся линейностью краевого условия (17.3). Имеем:

$$l_a[y] = l_a[Cu + v] = Cl_a[u] + l_a[v] = A \qquad \forall C \in \mathrm{R}$$

$$\Leftrightarrow \begin{cases} l_a[u] = 0, \\ l_a[v] = A \end{cases} \Leftrightarrow \begin{cases} \alpha_0 u(a) + \alpha_1 u'(a) = 0, & (17.11) \\ \alpha_0 v(a) + \alpha_1 v'(a) = A. & (17.12) \end{cases}$$

Равенство (17.11) будет выполнено, если положить, например,

$$u(a) = \alpha_1, \quad u'(a) = -\alpha_0. \tag{17.13}$$

Для удовлетворения равенства (17.12) можно взять

$$v(a) = \frac{A}{\alpha_0}, \quad v'(a) = 0, \tag{17.14}$$

если $\alpha_0 \neq 0$, или

$$v(a) = 0, \quad v'(a) = \frac{A}{\alpha_1}, \tag{17.15}$$

если $\alpha_1 \neq 0$. Заметим, что одновременно α_0 и α_1 в нуль не обращаются, в силу условия (17.5); неоднозначность задания начальных условий для u и v связана с лишней степенью свободы: одна неизвестная функция y заменяется линейной комбинацией двух функций (u и v).

Итак, согласно (17.10), функции u и v в представлении решения (17.8) можно найти, решая уравнения

$$u'' + p(x)u' + q(x)u = 0 \tag{17.16}$$

и

$$v'' + p(x)v' + q(x)v = f(x) \tag{17.17}$$

с начальными условиями (17.13) для (17.16) и с начальными условиями (17.14) или (17.15) для (17.17), применяя какие-либо методы решения задач Коши для уравнений второго порядка (см. гл.15). Приближенное решение этих задач строится на отрезке $[a, b]$, в результате чего становятся известными, в частности, значения $u(b)$, $u'(b)$, $v(b)$, $v'(b)$. Это позволяет подобрать постоянную C так, чтобы с этим значением C и найденными функциями $u(x)$ и $v(x)$ функция $y = Cu + v$ удовлетворяла не только уравнению (17.2) и условию (17.3), но и второму условию (17.4). Имеем:

$$l_b[y] = l_b[Cu + v] = Cl_b[u] + l_b[v] = B,$$

если

$$C = \frac{B - l_b[v]}{l_b[u]} = \frac{B - \beta_0 v(b) - \beta_1 v'(b)}{\beta_0 u(b) + \beta_1 u'(b)}. \tag{17.18}$$

Замечание 17.1. Если знаменатель выражения C в (17.18) окажется равным нулю, это будет означать, что однородная краевая задача

$$L[u] = 0, \quad l_a[u] = 0, \quad l_b[u] = 0$$

имеет нетривиальное решение $u(x)$, что, в свою очередь, служит признаком вырожденности исходной задачи (17.2)–(17.4).

Метод дифференциальной прогонки*). При выводе метода прогонки решения линейных алгебраических систем с трехдиагональными матрицами коэффициентов, т.е. для трехточечных разностных уравнений второго порядка, связывающих три соседние компоненты вектора неизвестных системы, в гл. 2 мы исходили из предположения о наличии связей между каждыми двумя соседними неизвестными и получали рекуррентные формулы для коэффициентов таких связей. Проводя параллель между линейными дифференциальными и линейными разностными

*) Метод разработан И.М.Гельфандом и О.В.Локуциевским [20, 131].

уравнениями, предположим, что существуют такие две функции $\delta = \delta(x)$ и $\gamma = \gamma(x)$, с помощью которых решение $y = y(x)$ данного линейного дифференциального уравнения второго порядка (17.2) может быть представлено как решение уравнения первого порядка

$$y' = \delta(x)y + \gamma(x). \qquad (17.19)$$

Продифференцируем это равенство и подставим выражение второй производной $y'' = \delta' y + \delta y' + \gamma'$ в исходное уравнение (17.2). Имеем

$$\delta' y + \delta y' + \gamma' + py' + qy = f,$$

откуда, выразив y', получаем равенство типа (17.19):

$$y' = -\frac{\delta' + q}{\delta + p} y + \frac{f - \gamma'}{\delta + p}. \qquad (17.20)$$

Уравнения (17.20) и (17.19) можно отождествлять при условии, что

$$\begin{cases} \delta = -\dfrac{\delta' + q}{\delta + p}, \\ \gamma = \dfrac{f - \gamma'}{\delta + p}. \end{cases}$$

Следовательно, функции $\delta = \delta(x)$ и $\gamma = \gamma(x)$ должны удовлетворять системе дифференциальных уравнений первого порядка

$$\begin{cases} \delta' = -p(x)\delta - \delta^2 - q(x), & (17.21) \\ \gamma' = -(p(x) + \delta)\gamma + f(x). & (17.22) \end{cases}$$

Чтобы найти начальные условия для уравнений системы (17.21), (17.22), учтем, что решение $y(x)$ в точке a должно удовлетворять уравнению (17.19), т.е. должно выполняться равенство

$$y'(a) = \delta(a)y(a) + \gamma(a). \qquad (17.23)$$

С другой стороны , в предположении, что $\alpha_1 \neq 0$, из краевого условия (17.3) для того же решения в точке a имеем

$$y'(a) = -\frac{\alpha_0}{\alpha_1} y(a) + \frac{A}{\alpha_1}. \qquad (17.24)$$

Сравнение (17.23) с (17.24) приводит к начальным условиям

$$\begin{cases} \delta(a) = -\dfrac{\alpha_0}{\alpha_1}, & (17.25) \\ \gamma(a) = \dfrac{A}{\alpha_1}. & (17.26) \end{cases}$$

Решив на отрезке $[a, b]$ задачу Коши для системы (17.21), (17.22) с начальными условиями (17.25), (17.26), получим требуемые для уравнения (17.19) функции $\delta(x)$ и $\gamma(x)$. Для нахождения частного решения $y(x)$ уравнения (17.19), служащего одновременно решением данной краевой задачи, аналогично предыдущему запишем равенство

$$y'(b) = \delta(b)y(b) + \gamma(b)$$

и подставим это выражение $y'(b)$ в краевое условие (17.4):

$$\beta_0 y(b) + \beta_1 \delta(b) y(b) + \beta_1 \gamma(b) = B.$$

Отсюда получаем привязку решения $y(x)$ в точке b:

$$y(b) = \frac{B - \beta_1 \gamma(b)}{\beta_0 + \beta_1 \delta(b)} \qquad (17.27)$$

(если $\beta_0 + \beta_1 \delta(b) \neq 0$). Остается решить задачу Коши (17.19), (17.27) от точки b к точке a.

Таким образом, ***метод дифференциальной прогонки*** *решения линейной краевой задачи* (17.2)–(17.4) *заключается в решении трех начальных задач для дифференциальных уравнений первого порядка: сначала параллельно (или последовательно в записанном порядке) решаются уравнения* (17.21), (17.22) *с начальными условиями* (17.25), (17.26) *в прямом направлении от* a *к* b*, затем в обратном направлении от* b *к* a *решается уравнение* (17.19) *с начальным условием* (17.27).

При численной реализации этого метода следует обратить внимание на необходимость согласования расчетных сеток решения начальных задач в одном и в другом направлениях так, чтобы имелась возможность получения решения последнего уравнения (17.19) с нужной точностью при условии, что в процессе численного интегрирования вспомогательных уравнений (17.21), (17.22) находятся лишь каркасы функций $\delta(x)$ и $\gamma(x)$ на своих сетках.

Замечание 17.2. Если не выполняется сделанное выше предположение $\alpha_1 \neq 0$, то вместо описанной дифференциальной прогонки, которую естественно назвать ***левой***, можно применить ***правую*** прогонку,

предполагая $\beta_1 \neq 0$. Для этого достаточно начальные условия для нахождения функций $\delta(x)$ и $\gamma(x)$ вывести из краевого условия (17.4), а для $y(x)$ — из (17.3). В результате придется строить вспомогательные решения, т.е. решать систему (17.21), (17.22) в направлении от b к a, а окончательное решение $y(x)$ находить из уравнения (17.19) в прямом направлении от a к b. Очевидно, что когда $\alpha_1 = \beta_1 = 0$, т.е. в случае первой краевой задачи, рассмотренный выше метод дифференциальной прогонки неприменим (можно сказать, что они дополняют друг друга с методом пристрелки).

Имеются и другие способы сведения краевых задач к начальным (см., например, [92]).

17.3. МЕТОД КОНЕЧНЫХ РАЗНОСТЕЙ

Идея ***метода конечных разностей*** (МКР*)) решения краевых задач весьма проста и видна уже из самого названия: вместо производных в дифференциальном уравнении используются их конечноразностные аппроксимации. Эта идея уже применялась нами в § 16.2 при построении простейших разностных схем решения задачи Коши. В отличие от того случая, где дискретным аналогом начальной дифференциальной задачи служила начальная задача для разностного уравнения, что позволяло последовательно, шаг за шагом, находить компоненты каркаса приближенного решения, здесь, т.е. для краевой дифференциальной задачи, будет получаться краевая же задача для разностного уравнения, а это ведет к необходимости решать алгебраическую систему. Ясно, что при построении дискретных аппроксимаций краевых дифференциальных задач нужно стремиться увязать две, возможно, противоречивые цели: хорошее качество аппроксимации и эффективное и устойчивое решение получающихся при этом алгебраических систем.

Рассмотрим наиболее типичные воплощения указанной идеи МКР для линейной краевой задачи (17.2)–(17.4).

Сначала вводим на отрезке $[a, b]$ ***сетку*** с шагом $h = \dfrac{b-a}{n}$:

$$\omega_h := \{x_i \mid x_i = x_0 + ih; \quad i = 0, 1, \ldots, n; \quad x_0 := a,\, x_n := b\}.$$

На этой сетке определяются ***сеточные функции***

$$p_i := p(x_i), \quad q_i := q(x_i), \quad f_i := f(x_i), \tag{17.28}$$

отвечающие функциональным коэффициентам данного дифференциального уравнения (17.2). Считая $y(x)$ точным решением

*) Зарубежная аббревиатура FDM — от англ. *finite difference method.*

данной краевой задачи (17.2)–(17.4), через

$$y_i \approx y(x_i) \tag{17.29}$$

будем обозначать i-ю компоненту искомого каркаса приближенного решения $y_n(x) \approx y(x)$.

Фиксируя в уравнении (17.2) $x = x_i$, с учетом обозначений (17.28) приходим к равенствам

$$y''(x_i) + p_i y'(x_i) + q_i y(x_i) = f_i, \tag{17.30}$$

где целая переменная i может принимать значения от 0 до n по числу узлов сетки, а под $y(x_i)$, $y'(x_i)$, $y''(x_i)$ понимаются значения точного решения $y(x)$ и его производных в i-ом узле. В каждом внутреннем узле сетки ω_h, т.е. при $i = 1, 2, \ldots, n-1$, значения производных аппроксимируем конечноразностными отношениями по симметричным формулам второго порядка точности (см. (13.18) и (13.20)):

$$y'(x_i) = \frac{y(x_{i+1}) - y(x_{i-1})}{2h} + O(h^2),$$

$$y''(x_i) = \frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2} + O(h^2).$$

В результате подстановки последних в равенства (17.30) при $i = 1, 2, \ldots, n-1$ получаем

$$\frac{y(x_{i+1}) - 2y(x_i) + y(x_{i-1})}{h^2} + p_i \frac{y(x_{i+1}) - y(x_{i-1})}{2h} + q_i y(x_i) = f_i + O(h^2).$$

Это уравнение служит точным отражением дифференциальной связи (17.2), но в нем имеется неопределенное слагаемое $O(h^2)$. Отбрасывая его, приходим к разностному уравнению относительно приближенных значений решения (обозначения которых соответствуют (17.29)):

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + p_i \frac{y_{i+1} - y_{i-1}}{2h} + q_i y_i = f_i. \tag{17.31}$$

После приведения подобных членов в (17.31) получаем стандартное трехточечное разностное уравнение второго порядка

$$\left(1 + \frac{h}{2} p_i\right) y_{i+1} - (2 - h^2 q_i) y_i + \left(1 - \frac{h}{2} p_i\right) y_{i-1} = h^2 f_i, \tag{17.32}$$

где $i = 1, 2, \ldots, n-1$.

Интерпретируя (17.32) как компактную запись системы линейных алгебраических уравнений с трехдиагональной матрицей коэффициентов, видим, что число уравнений в ней $n-1$, в то время как неизвестных — $n+1$: $y_0, y_1, \ldots, y_n$. Два недостающие уравнения этой системы (или, в другой терминологии, краевые условия разностного уравнения (17.32)) следует получить на основе краевых условий (17.3), (17.4) данной задачи.

Будем рассматривать два варианта аппроксимации входящих в краевые условия значений первой производной решения в точках $a = x_0$ и $b = x_n$. Именно, здесь идет речь о несимметричных аппроксимациях первого и второго порядков точности (достоинства и недостатки каждого из этих вариантов будут обсуждены позже).

Согласно формулам (13.15), (13.26), (13.14) и (13.27), можно записать равенства

$$y'(a) = \frac{y(x_1) - y(x_0)}{h} + O(h) = \frac{-3y(x_0) + 4y(x_1) - y(x_2)}{2h} + O(h^2),$$

$$y'(b) = \frac{y(x_n) - y(x_{n-1})}{h} + O(h) = \frac{y(x_{n-2}) - 4y(x_{n-1}) + 3y(x_n)}{2h} + O(h^2).$$

В первом варианте, при аппроксимации $y'(a)$ и $y'(b)$ в (17.3) и (17.4) двухточечными разностными отношениями первого порядка, имеем

$$\alpha_0 y(x_0) + \alpha_1 \frac{y(x_1) - y(x_0)}{h} + O(h) = A,$$

$$\beta_0 y(x_n) + \beta_1 \frac{y(x_n) - y(x_{n-1})}{h} + O(h) = B.$$

Отсюда после отбрасывания слагаемого $O(h)$ (с заменой $y(x_i)$ на y_i) и упрощения получаем краевые условия для разностного уравнения (17.32), иначе, нулевое и n-е уравнения

$$(h\alpha_0 - \alpha_1)y_0 + \alpha_1 y_1 = Ah \qquad (17.33)$$

и

$$-\beta_1 y_{n-1} + (h\beta_0 + \beta_1)y_n = Bh \qquad (17.34)$$

системы алгебраических уравнений, задаваемой равенством (17.32).

Во втором варианте имеем аналогично

$$\alpha_0 y(x_0) + \alpha_1 \frac{-3y(x_0) + 4y(x_1) - y(x_2)}{2h} + O(h^2) = A,$$

$$\beta_0 y(x_n) + \beta_1 \frac{y(x_{n-2}) - 4y(x_{n-1}) + 3y(x_n)}{2h} + O(h^2) = B,$$

откуда следуют дополнительные к (17.32) связи между тремя первыми неизвестными

$$(2h\alpha_0 - 3\alpha_1)y_0 + 4\alpha_1 y_1 - \alpha_1 y_2 = 2Ah \qquad (17.35)$$

и тремя последними

$$\beta_1 y_{n-2} - 4\beta_1 y_{n-1} + (2h\beta_0 + 3\beta_1)y_n = 2Bh. \qquad (17.36)$$

Сравним теперь два рассматриваемых варианта. В первом из них СЛАУ, образованная уравнениями (17.33), (17.32), (17.34), имеет трехдиагональную матрицу коэффициентов, и к ней сразу можно применить высокоэффективный метод прогонки (§ 2.6). Во втором случае соответствующая методу прогонки структура матрицы коэффициентов СЛАУ еще должна быть создана, для чего нужно из уравнения (17.35) с помощью уравнения, получающегося из (17.32) при $i = 1$, исключить неизвестное y_2, а из (17.36) с помощью (17.32) при $i = n-1$ исключить y_{n-2}. Усложнения, сопутствующие второму варианту, могут быть оправданы тем, что в этом случае исходная дифференциальная краевая задача полностью аппроксимируется алгебраической системой относительно компонент каркаса решения с точностью $O(h^2)$, в то время как о первом варианте такого сказать нельзя (если, конечно, речь не идет о первой краевой задаче, т.е. о случае $\alpha_1 = \beta_1 = 0$, когда в аппроксимации краевых условий вообще нет нужды). Однако в поисках компромисса между качеством аппроксимации и численной устойчивостью при решении конкретных задач первый вариант может оказаться и более предпочтительным.

Остановимся теперь на вопросе **устойчивости** построенной конечноразностной схемы решения краевой задачи (17.2)–(17.4). Как отмечалось ранее, эту устойчивость можно связать с устойчивостью метода прогонки, что, в свою очередь, можно гарантировать, когда матрица коэффициентов имеет свойство диагонального преобладания (теорема 2.2). Посмотрим с этой точки зрения на i-е «внутреннее» уравнение системы, т.е. на уравнение (17.32).

Условие диагонального преобладания для (17.32) означает, что должно выполняться неравенство

$$\left|2 - h^2 q_i\right| > \left|1 + \frac{h}{2} p_i\right| + \left|1 - \frac{h}{2} p_i\right| \qquad \forall\, i \in \{1, 2, \ldots, n-1\}. \qquad (17.37)$$

Рассмотрим, что представляет собой правая часть этого неравенства. Раскрывая модули, имеем

$$\left|1 + \frac{h}{2} p_i\right| + \left|1 - \frac{h}{2} p_i\right| = \begin{cases} -hp_i, & \text{если } hp_i < 2, \\ 2, & \text{если } |hp_i| \le 2, \\ hp_i, & \text{если } hp_i > 2. \end{cases}$$

Следовательно, правую часть неравенства (17.37) как функцию переменной hp_i (считая ее изменяющейся непрерывно) в условных координатах можно представить в виде графика, изображенного на рис.17.2.

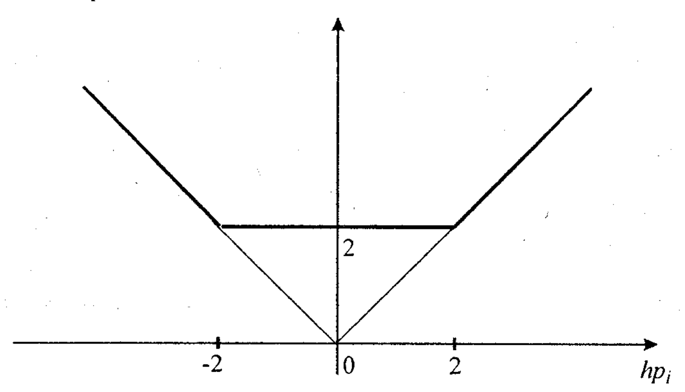


Рис. 17.2. Условный график правой части неравенства (17.37)

Так как левая часть неравенства (17.37) при $q_i > 0$ и малых $h > 0$ (малость h нужна из требований аппроксимации) меньше 2, то на устойчивость прогонки можно рассчитывать лишь в случае, когда $q(x) < 0$. При этом имеет место

$$\left|2 - h^2 q_i\right| = 2 - h^2 q_i > 2 \quad \forall\, h.$$

Чтобы в таком случае неравенство (17.37) выполнялось при любых $p(x)$, для правой его части считаем допустимым только значение 2 (т.е. используем горизонтальную часть графика на рис.17.2). Отсюда получаем ограничение

$$|hp_i| \le 2,$$

означающее, что устойчивость прогонки можно гарантировать при условии, что шаг дискретизации h удовлетворяет неравенству

$$h \le \frac{2}{|p_i|} \quad \forall\, i \in \{1, 2, \ldots, n-1\}.$$

Усиливая это неравенство и используя сформулированное в конце § 16.1 утверждение «Аппроксимация плюс устойчивость дает сходимость», приходим к заключению, что *если в дифференциальном уравнении* (17.2)

$$q(x) < 0 \quad \forall\, x \in [a, b], \tag{17.38}$$

а в определяющем МКР разностном уравнении (17.32)

$$h \le \frac{2}{\max\limits_{x\in[a,b]} |p(x)|},$$

то МКР *с х о д и т с я* (по крайней мере, к решению первой краевой задачи, т.е. когда в (17.3), (17.4) $\alpha_1 = \beta_1 = 0$; в других случаях требуется более детальный анализ).

Наличие ограничения на шаг h в методе конечных разностей второго порядка (17.32) характеризует его как условно устойчивый метод. Если отказаться от аппроксимации всех производных с порядком $O(h^2)$ и использовать в роли $y'(x_i)$ правые или левые разностные отношения первого порядка точности, связывая их выбор со знаком p_i, а именно, рассматривая вместо (17.31) разностное уравнение

$$\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} + p_i \begin{bmatrix} \dfrac{y_{i+1} - y_i}{h}, \text{ если } & p_i > 0 \\ \dfrac{y_i - y_{i-1}}{h}, \text{ если } & p_i < 0 \end{bmatrix} + q_i y_i = f_i$$

при $i = 1, 2, \ldots, n-1$, придем к конечноразностному методу

$$\begin{cases} y_{i-1} - (2 + hp_i - h^2 q_i) y_i + (1 + hp_i) y_{i+1} = h^2 f_i, & \text{если } p_i > 0, \\ (1 - hp_i) y_{i-1} - (2 - hp_i - h^2 q_i) y_i + y_{i+1} = h^2 f_i, & \text{если } p_i < 0 \end{cases}, \tag{17.39}$$

имеющему первый порядок точности независимо от точности аппроксимации краевых условий.

Легко видеть, что при условии (17.38) диагональное преобладание в методе (17.39) будет при любой величине шага $h > 0$. Отсюда следует его *безусловная устойчивость*, правда в ущерб точности; последнее означает необходимость проведения вычислений с более мелким шагом для доведения погрешности решения до некоторой фиксированной малой величины, чем это требует метод второго порядка (17.32), если они оба одновременно применимы. Конечноразностный метод (17.39) широко используется при решении задач динамики жидкости и газов и называется ***upwind***- или ***противопотоковым методом***.

17.4. МЕТОД КОЛЛОКАЦИИ

Будем искать приближенное решение линейной краевой задачи (17.2)–(17.4) в виде функции

$$y_n(x) := \varphi_0(x) + \sum_{i=1}^{n} c_i \varphi_i(x), \tag{17.40}$$

где определяемые на отрезке $[a,b]$ ***базисные функции*** $\varphi_i(x)$ ($i = 1, 2, \ldots, n$) и дополнительная функция $\varphi_0(x)$ должны быть дважды дифференцируемыми и попарно линейно независимыми. Кроме того, функция $\varphi_0(x)$ должна удовлетворять данным краевым условиям (17.3), (17.4), а функции $\varphi_i(x)$ при $i = 1, 2, \ldots, n$ — соответствующим однородным краевым условиям, т.е. должны выполняться равенства

$$\left.\begin{aligned} \alpha_0\varphi_i(a) + \alpha_1\varphi_i'(a) = 0\,, \\ \beta_0\varphi_i(b) + \beta_1\varphi_i'(b) = 0 \end{aligned}\right\} \quad \forall\, i \in \{\, i = 1, 2, \ldots, n\,\}. \qquad (17.41)$$

В таком случае функция $y_n(x)$, определяемая выражением (17.40), при любых значениях коэффициентов c_i гарантированно удовлетворяет краевым условиям (17.3), (17.4). Действительно, например, в точке $x = a$ имеем

$$\alpha_0 y_n(a) + \alpha_1 y_n'(a) =$$

$$= \alpha_0\varphi_0(a) + \alpha_1\varphi_0'(a) + \alpha_0\sum_{i=1}^{n} c_i\varphi_i(a) + \alpha_1\sum_{i=1}^{n} c_i\varphi_i'(a) =$$

$$= A + \sum_{i=1}^{n} c_i\left[\alpha_0\varphi_i(a) + \alpha_1\varphi_i'(a)\right] = A\,;$$

аналогично при $x = b$ с помощью (17.4), (17.40) и (17.41) проверяется справедливость равенства

$$\beta_0 y_n(b) + \beta_1 y_n'(b) = B\,.$$

Представление приближенного решения $y_n(x)$, подобное (17.40), характерно для многих приближенно-аналитических методов решения краевых задач (возможны вариации требований к базисным функциям); главное их различие состоит в том, на какой основе ищутся коэффициенты c_i в линейной комбинации базисных функций $\varphi_i(x)$ выражения (17.40).

В ***методе коллокации****) коэффициенты c_i в представлении (17.40) приближенного решения $y_n(x)$ подбираются так, чтобы в ***узлах коллокации*** x_i таких, что

$$a < x_1 < x_2 < \ldots < x_n < b$$

(не обязательно равноотстоящих, но строго внутренних точках отрезка $[a, b]$), *значения* $y_n(x_i)$ *приближенного решения были согласованы с точными значениями* $y(x_i)$.

---

*) *Collocatio* (лат.) — размещение, расстановка. В книге [80] такой метод называют ***интерполяционным методом*** или ***методом совпадений***.

Поскольку точное решение $y(x)$ задачи (17.2)–(17.4) неизвестно, согласование $y_n(x)$ и $y(x)$ в узлах коллокации x_i проводим подстановкой $y_n(x)$ в уравнение (17.2). Имеем равенство

$$y_n''(x_i) + p(x_i)y_n'(x_i) + q(x_i)y_n(x_i) = f(x_i)\,, \qquad (17.42)$$

которое, в силу выставляемого требования согласования $y_n(x_i)$ с $y(x_i)$, считаем точным при каждом $i \in \{1, 2, \ldots, n\}$. Продифференцировав дважды функцию $y_n(x)$ в представлении (17.40), от равенства (17.42) переходим к равенству

$$\sum_{j=1}^{n} c_j\varphi_j''(x_i) + p_i\sum_{j=1}^{n} c_j\varphi_j'(x_i) + q_i\sum_{j=1}^{n} c_j\varphi_j(x_i) =$$

$$= f_i - \varphi_0''(x_i) - p_i\varphi_0'(x_i) - q_i\varphi_0(x_i)\,, \qquad (17.43)$$

где p_i, q_i, f_i соответствуют обозначениям (17.28) сеточных функций. Положим

$$a_{ij} := \varphi_j''(x_i) + p_i\varphi_j'(x_i) + q_i\varphi_j(x_i)\,, \qquad (17.44)$$

$$b_i := f_i - \varphi_0''(x_i) - p_i\varphi_0'(x_i) - q_i\varphi_0(x_i)\,. \qquad (17.45)$$

Тогда (17.43) приобретает стандартный вид линейной алгебраической системы

$$\sum_{j=1}^{n} a_{ij}c_j = b_i, \quad i = 1, 2, \ldots, n \qquad (17.46)$$

относительно коэффициентов $c_1, c_2, \ldots, c_n$. Решив эту систему каким-нибудь стандартным методом и подставив найденные коэффициенты c_i в выражение (17.40), получаем приближенное решение $y_n(x)$.

Успех применения метода коллокации к задаче (17.2)–(17.4), впрочем, как и других приближенно-аналитических методов, сильно зависит от удачного выбора базисных функций $\varphi_i(x)$ в представлении приближенного решения (17.40). В конкретных задачах выбор таких функций, по возможности, должен опираться на априорные или эмпирические сведения о решении. В отсутствие таковых, т.е. в рассматриваемом абстрактном случае, для смешанной краевой задачи (17.2)–(17.4) можно предложить, например, следующий **набор базисных функций**.

В качестве φ_0 возьмем линейную функцию

$$\varphi_0(x) = \delta + \gamma x\,, \qquad (17.47)$$

коэффициенты которой подберем так, чтобы она удовлетворяла неоднородным краевым условиям (17.3), (17.4), т.е. из линейной

алгебраической системы

$$\begin{cases}\alpha_0\delta+(\alpha_0 a+\alpha_1)\gamma=A,\\ \beta_0\delta+(\beta_0 b+\beta_1)\gamma=B.\end{cases} \qquad (17.48)$$

Функции $\varphi_i(x)$ при $i=1,2,\dots,n$ можно взять однопараметрическими вида

$$\varphi_i(x)=\gamma_i(x-a)^i+(x-a)^{i+1}, \qquad (17.49)$$

если в (17.3) $\alpha_1=0$, или вида

$$\varphi_i(x)=\gamma_i(x-a)^{i+1}+(x-a)^{i+2} \qquad (17.50)$$

в самом общем случае. Очевидно, что при любых γ_i эти функции удовлетворяют первому из требуемых равенств (17.41)*), а если зафиксировать

$$\gamma_i=-\frac{\beta_0(b-a)^2+(i+1)\beta_1(b-a)}{\beta_0(b-a)+i\beta_1} \qquad (17.51)$$

в выражении (17.49) и

$$\gamma_i=-\frac{\beta_0(b-a)^2+(i+2)\beta_1(b-a)}{\beta_0(b-a)+(i+1)\beta_1} \qquad (17.52)$$

в (17.50), то они будут подчиняться и второму из этих равенств. Следовательно, можно рассчитывать, что с такими базисными функциями при найденных методом коллокации (или каким-либо другим методом) коэффициентах c_i определенная посредством (17.40) функция $y_n(x)$ будет удовлетворять краевым условиям и может служить приближенным решением данной краевой задачи (17.2)–(17.4).

Проблема формального выбора базисных функций φ_i значительно упрощается в случае, когда в задаче (17.2)–(17.4) фигурируют однородные краевые условия первого рода, т.е. когда

$$y(a)=0, \quad y(b)=0. \qquad (17.53)$$

В такой ситуации в выражении (17.40) не нужна функция φ_0, а в роли φ_i $(i=1,2,\dots,n)$ могут выступать, например, функции

$$\varphi_i(x)=(x-a)^i(b-x)$$

*) Если $\alpha_1\neq 0$, то первому из равенств (17.41) не удовлетворяет одна функция семейства (17.49), а именно $\varphi_1(x)$.

или

$$\varphi_i(x)=\sin\frac{i(x-a)}{b-a}\pi.$$

К этому случаю, т.е. к условиям (17.53), легко свести более общий случай неоднородных краевых условий первого рода

$$y(a)=A, \quad y(b)=B. \qquad (17.54)$$

С этой целью достаточно сделать линейную замену (линейный сдвиг)

$$y=u+v, \quad \text{где} \quad v:=A+\frac{B-A}{b-a}(x-a).$$

Дважды дифференцируя эту функцию y и подставляя результаты в уравнение (17.2), от задачи (17.2), (17.54) приходим к краевой задаче с однородными краевыми условиями относительно новой переменной u:

$$u''+p(x)u'+q(x)u=f(x)-\frac{B-A}{b-a}p(x)-vq(x), \qquad x\in[a,b],$$
$$u(a)=0, \quad u(b)=0.$$

Пример 17.1. Применим метод коллокации для приближенного представления решения краевой задачи

$$\begin{aligned}&x^4y''+x^6y'-x^5y=6-3x^3, \quad x\in[1,2],\\ &y(1)=1, \quad 3y(2)+y'(2)=0.5.\end{aligned} \qquad (17.55)$$

Составив систему (17.48) относительно коэффициентов δ и γ функции $\varphi_0(x)$ вида (17.47)

$$\begin{cases}\delta+\gamma=1,\\ 3\delta+7\gamma=0.5,\end{cases}$$

находим удовлетворяющую данным краевым условиям линейную функцию $\varphi_0(x)=\frac{13}{8}-\frac{5}{8}x$. Ограничиваясь одной базисной функцией $\varphi_1(x)$ вида (17.49), по формуле (17.51) вычисляем коэффициент

$$\gamma_1=-\frac{3(2-1)^2+2(2-1)}{3(2-1)+1\cdot 1}=-\frac{5}{4},$$

подстановка которого в (17.49) при $i=1$ дает базисную функцию

$$\varphi_1(x)=(x-1)^2-\frac{5}{4}(x-1).$$

Возьмем в качестве узла коллокации середину рассматриваемого промежутка — точку $x_1=\frac{3}{2}$ — и потребуем, чтобы функция

$$y_1(x)=\varphi_0(x)+c_1\varphi_1(x)$$

удовлетворяла в этой точке заданному дифференциальному уравнению. Подставив в него

$$y_1\left(\frac{3}{2}\right)=\frac{11}{16}-\frac{3}{8}c_1, \quad y_1'=-\frac{5}{8}-\frac{1}{4}c_1 \quad \text{и} \quad y_1''=2c_1,$$

получаем $c_1=\frac{701}{864}$. Таким образом, простейшая коллокация с одним узлом приводит к приближению решения данной краевой задачи (17.55) квадратичной функцией

$$y_1(x)=1-\frac{5}{8}(x-1)+\frac{701}{864}(x-1)(x-\frac{9}{4})$$

(ее точное решение $y(x)=\frac{1}{x^2}$).

Замечание 17.3. Приближенное решение $y_n(x)$ (17.40), согласно идее метода коллокации, в точках коллокации $x_i\in(a,\,b)$ должно удовлетворять данному дифференциальному уравнению (17.2). Казалось бы, в этих точках оно должно совпадать с точным решением $y(x)$ данной краевой задачи (17.2)–(17.4). Однако это не так, в чем легко убедиться, сравнив полученное в примере 17.1 приближенное решение в узле коллокации $x_1=\frac{3}{2}$, равное $y_1\left(\frac{3}{2}\right)\approx 0.38$, с точным значением решения $y\left(\frac{3}{2}\right)=0.(4)$ (см. также сравнение графиков $y_1(x)$ и $y(x)$ далее на рис.17.3)

Замечание 17.4. Нет принципиальных препятствий к привлечению метода коллокации для приближенного решения н е л и н е й н ы х краевых задач. Трудности на этом пути ожидают при подборе базисных функций φ_i, удовлетворяющих краевым условиям (17.1) в случае их нелинейности, и в необходимости решать нелинейные системы при отыскании коэффициентов c_i того же представления (17.40) приближенного решения. Например, если данное дифференциальное уравнение имеет вид

$$y''=f(x,\,y,\,y'),$$

то коэффициенты $c_1, c_2, \ldots, c_n$ в представлении $y_n(x)$ вида (17.40) должны находиться из системы

$$\sum_{j=0}^{n}c_j\varphi_j''(x_i)=f(x_i,\,\sum_{j=0}^{n}c_j\varphi_j(x_i),\,\sum_{j=0}^{n}c_j\varphi_j'(x_i)),$$

где $i=1,2,\ldots,n$, а $c_0:=1$ (или $c_0:=0$, если это дифференциальное уравнение сопровождается краевыми условиями первого рода).

17.5. МЕТОД ГАЛЁРКИНА

Чтобы лучше понять основную идею ***проекционных методов***, наиболее ярким представителем которых является метод Галёркина, ненадолго отвлечемся от рассматриваемой краевой задачи и сделаем небольшой экскурс в функциональный анализ.

Пусть L — некоторый л и н е й н ы й оператор, действующий в ***гильбертовом пространстве*** H, т.е. в полном нормированном пространстве со скалярным произведением $(\cdot,\cdot)$. Стоит задача приближенного решения ***операторного уравнения***

$$Ly=f, \tag{17.56}$$

т.е. задача отыскания некоторого приближения к неизвестному элементу $y\in H$, соответствующему заданному элементу $f\in H$. Пусть, далее, $\{\varphi_i\}_{i=1}^{\infty}$ — некоторая полная замкнутая система линейно независимых элементов из H. Ее n первых элементов $\varphi_1,\ldots,\varphi_n$ выделяют в H конечномерное подпространство H_n, в котором и ищется приближенное решение уравнения (10.56):

$$y_n:=\sum_{i=1}^{n}c_i\varphi_i. \tag{17.57}$$

Для удобства будем считать, что элемент Ly_n принадлежит тому же подпространству H_n. Тогда к тривиальному равенству

$$f=Ly_n+(f-Ly_n)$$

можно применить одну из центральных теорем теории гильбертовых пространств, согласно которой *любой элемент гильбертова пространства может быть представлен в виде суммы определенного элемента подпространства* (проекции данного элемента на подпространство) *и определенного элемента пространства, ортогонального к выбранному подпространству* (реализующего расстояние от исходного элемента до его проекции) [45, 148]*).

Принадлежность элемента $f-Ly_n$ ортогональному к H_n подпространству $H_n^{\perp}$ означает, что он ортогонален каждому элементу φ_i, входящему в базис пространства H_n. Таким обра-

*) Представьте себе, например, точку трехмерного пространства в декартовой системе координат, спроектируйте ее радиус-вектор на одну из координатных плоскостей; полученный при этом прямоугольный векторный треугольник может служить простейшей геометрической моделью этой теоремы.

зом, при любом $i = 1, 2, \dots, n$ имеем:

$$(f - Ly_n) \perp \varphi_i \quad \Leftrightarrow \quad (f - Ly_n, \varphi_i) = 0 \quad \Leftrightarrow \quad (Ly_n, \varphi_i) = (f, \varphi_i).$$

Подставляя сюда выражение y_n (10.57) и пользуясь простейшими свойствами скалярного произведения, получаем:

$$\left(L \sum_{j=1}^{n} c_j \varphi_j, \varphi_i \right) = (f, \varphi_i) \quad \Leftrightarrow \tag{17.58}$$

$$\Leftrightarrow \quad \sum_{j=1}^{n} (L\varphi_j, \varphi_i) c_j = (f, \varphi_i).$$

Итак, ***метод Галёркина****)* *приближенного решения* о п е р а т о р н о г о *уравнения* (10.56) *сводится к нахождению коэффициентов* $c_1, \dots, c_n$ *линейной комбинации некоторых задаваемых определенным образом линейно независимых функций* $\varphi_1, \dots, \varphi_n$ (*называемых* ***базисными*** *или* ***координатными функциями***) *так, чтобы эти коэффициенты удовлетворяли линейной системе*

$$\sum_{j=1}^{n} a_{ij} c_j = d_i, \quad i = 1, 2, \dots, n, \tag{17.59}$$

где

$$a_{ij} := (L\varphi_j, \varphi_i), \quad d_i := (f, \varphi_i). \tag{17.60}$$

Вернемся к краевой задаче (17.2)–(17.4). Естественным для ее рассмотрения пространством является гильбертово пространство $L_2[a, b]$ функций, интегрируемых на отрезке $[a, b]$ с квадратом. Скалярное произведение здесь определяется равенством

$$(u, v) = \int_a^b u(x) v(x) dx. \tag{17.61}$$

Как и в предыдущем параграфе, будем искать приближенное решение задачи (17.2)–(17.4) в виде задаваемой формулой (17.40) функции $y_n(x)$ такой, чтобы она удовлетворяла данным краевым условиям, а это будет, как мы уже знаем, если

$$l_a[\varphi_0] = A, \quad l_b[\varphi_0] = B,$$

а

$$l_a[\varphi_i] = 0, \quad l_b[\varphi_i] = 0 \quad \forall i \in \{1, 2, \dots, n\}.$$

*) Галёркин Борис Григорьевич (1871–1945) — русский инженер и ученый в области теории упругости.

Подставляя $y_n(x)$ в данное уравнение (17.2), в силу линейности определенного там дифференциального оператора $L[y]$, имеем:

$$L\left[\varphi_0 + \sum_{i=1}^{n} c_i \varphi_i\right] = f \quad \Leftrightarrow \quad L\left[\sum_{i=1}^{n} c_i \varphi_i\right] = f - L[\varphi_0].$$

Следовательно, в таком случае, наличие дополнительного слагаемого φ_0 в представлении приближенного решения (17.40) вместо представления (17.57) означает, что равенство (17.58) нужно переписать в виде

$$\left(L \sum_{j=1}^{n} c_j \varphi_j, \varphi_i \right) = (f - L\varphi_0, \varphi_i),$$

а это, в свою очередь, вносит поправку в выражение d_i, зафиксированное равенством (17.60). Согласно этому замечанию и определению (17.61) скалярного произведения, находим выражение свободного члена d_i i-го уравнения системы (17.59):

$$d_i = (f - L[\varphi_0], \varphi_i) =$$

$$= \int_a^b [f(x) - \varphi_0''(x) - p(x)\varphi_0'(x) - q(x)\varphi_0(x)] \varphi_i(x) dx. \tag{17.62}$$

Теперь будем выражать коэффициенты a_{ij} СЛАУ (17.59) в соответствии с записанной в (17.60) формулой и скалярным произведением (17.61):

$$a_{ij} = (L\varphi_j, \varphi_i) = \int_a^b [\varphi_j''(x) + p(x)\varphi_j'(x) + q(x)\varphi_j(x)] \varphi_i(x) dx =$$

$$= \int_a^b \varphi_j''(x)\varphi_i(x) dx + \int_a^b p(x)\varphi_j'(x)\varphi_i(x) dx + \int_a^b q(x)\varphi_j(x)\varphi_i(x) dx.$$

Первый из интегралов в этом выражении преобразуем «по частям»:

$$\int_a^b \varphi_j''(x)\varphi_i(x) dx = \varphi_i(x)\varphi_j'(x)\Big|_a^b - \int_a^b \varphi_j'(x)\varphi_i'(x) dx,$$

что приводит к следующей формуле для вычисления коэффициентов при неизвестных $c_1, c_2, \dots, c_n$ системы (17.59):

$$a_{ij} = \varphi_i(b)\varphi_j'(b) - \varphi_i(a)\varphi_j'(a) - \int_a^b \varphi_j'(x)\varphi_i'(x) dx +$$

$$+ \int_a^b p(x)\varphi_j'(x)\varphi_i(x) dx + \int_a^b q(x)\varphi_j(x)\varphi_i(x) dx. \tag{17.63}$$

Выполненное преобразование интеграла, во-первых, активизировало роль краевых условий при построении приближенного решения $y_n(x)$ в форме (17.40) (напомним, что функции $\varphi_i(x)$ при $i=1,2,\ldots,n$ должны удовлетворять однородным краевым условиям (17.41)), во-вторых, позволило ослабить требования к гладкости базисных функций $\varphi_i(x)$, поскольку теперь, как видно из (17.63), у них могут быть даже разрывы производных (скачки). Подмеченный факт говорит о том, что методом Галёркина при подходящем выборе базисных функций можно находить решения краевых задач, определенные в каком-либо обобщенном смысле.

Пример 17.2. Снова обратимся к конкретной краевой задаче (17.55), к решению которой $y(x)=\dfrac{1}{x^2}$ в примере 17.1 было найдено квадратичное приближение методом коллокации.

Используя такое же представление

$$\widetilde{y}_1(x)=\varphi_0(x)+\widetilde{c}_1\varphi_1(x)=\frac{13}{8}-\frac{5}{8}x+\widetilde{c}_1\left[(x-1)^2-\frac{5}{4}(x-1)\right], \qquad (17.64)$$

будем теперь искать его коэффициент $\widetilde{c}_1$ методом Галёркина. Приведя данное уравнение в соответствие с видом (17.2), согласно которому имеем

$$p(x)=x^2, \qquad q(x)=-x, \qquad f(x)=\frac{6}{x^4}-\frac{3}{x},$$

по формуле (17.63) находим

$$a_{11}=\left(1-\frac{5}{4}\right)\left(2-\frac{5}{4}\right)-\int_1^2\left[2(x-1)-\frac{5}{4}\right]^2dx+$$

$$+\int_1^2 x^2\left[2(x-1)-\frac{5}{4}\right]\cdot\left[(x-1)^2-\frac{5}{4}(x-1)\right]dx-$$

$$-\int_1^2 x\left[(x-1)^2-\frac{5}{4}(x-1)\right]^2dx\approx-0.770,$$

а по формуле (17.62)

$$d_1=\int_1^2\left(\frac{6}{x^4}-\frac{3}{x}+\frac{13}{8}x\right)\left[(x-1)^2-\frac{5}{4}(x-1)\right]dx\approx-0.548.$$

Следовательно, в (17.64)

$$\widetilde{c}_1=\frac{d_1}{a_{11}}\approx\frac{0.548}{0.770}\approx0.712,$$

т.е. галёркинское квадратичное приближение к точному решению

$$y(x)=\frac{1}{x^2} \text{ есть } \quad \widetilde{y}_1(x)\approx1-\frac{5}{8}(x-1)+0.712(x-1)\left(x-\frac{9}{4}\right).$$

В точке $x=\dfrac{3}{2}$ таким способом получаем несколько более точное приближение $\widetilde{y}_1\left(\dfrac{3}{2}\right)\approx0.42$ к $y\left(\dfrac{3}{2}\right)$, чем найденное ранее методом коллокации.

Графическое сравнение квадратичных приближений $y_1(x)$ и $\widetilde{y}_1(x)$ к решению $y(x)$ краевой задачи (17.55), полученных методом коллокации (см. пример 17.1) и методом Галёркина, представлено на рис. 17.3.

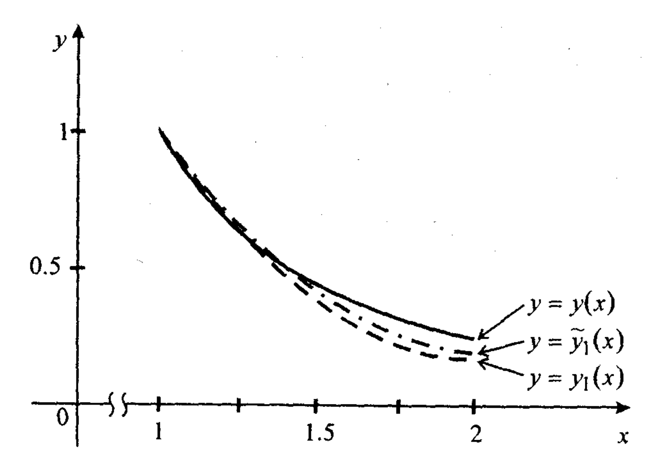


Рис. 17.3. Графики точного решения $y(x)$ краевой задачи (17.55) и его квадратичных приближений методами Галёркина ($\widetilde{y}_1(x)$) и коллокации ($y_1(x)$)

Замечание 17.5. Возможно, чаще в литературе, освещающей метод Галёркина для обыкновенных дифференциальных уравнений, исходное линейное уравнение второго порядка берут не в принятом здесь за основу виде (17.2), а в виде

$$-(k(x)y')'+q(x)y=f(x), \qquad (17.65)$$

имеющем прямую физическую интерпретацию и называемом в связи с этим ***одномерным стационарным уравнением теплопроводности*** или ***одномерным уравнением диффузии*** (см. [3, 4, 78, 94, 100, 138, 177 и др.]). При такой записи уравнения несколько упрощается выражение для подсчета коэффициентов a_{ij} системы (17.59) и, главное, становится более естественным анализ ситуации, когда коэффициенты в (17.65) могут оказаться разрывными.

Замечание 17.6. К той же системе (17.59) с коэффициентами (17.60) приходят и в ***методе Ритца***, опирающемся на вариационный принцип приближенного решения краевой задачи (17.2)–(17.4). Сведения об этом методе, а также другие подходы и варианты метода Галёркина можно почерпнуть в книгах [20, 81, 84, 117, 126, 147, 177 и др]. Схематично метод Ритца будет изложен далее в § 19.5.

17.6. МЕТОД КОНЕЧНЫХ ЭЛЕМЕНТОВ

При всех достоинствах метод Галёркина, рассмотренный в предыдущем параграфе, обладает тем существенным на современном этапе формализации прикладных исследований недостатком, что в «чистом» виде он малопригоден для автоматизированных компьютерных вычислений. Однако при подходящем выборе системы простых базисных функций φ_i в представлении приближенного решения $y_n(x)$, связанных с определенной на отрезке $[a, b]$ системой точек (сеткой), метод Галёркина трансформируется в сугубо численный процесс получения каркаса приближенного решения на заданной сетке, причем технология построения этого каркаса в конечном итоге оказывается близкой к той, которая присуща методу конечных разностей (см. § 17.3). Отсюда — название такого численного процесса ***проекционно-разностный*** или ***проекционно-сеточный метод***. Другое его более раннее и более употребительное название ***метод конечных элементов*** (МКЭ или FEM от англ. *finite element method*) связано с другими идеями, корни которых следует искать в строительной механике и теории упругости [135]. Преимущества этого метода перед многими другими методами обнаруживаются, в основном, при решении двумерных и трехмерных задач математической физики*), однако первые представления о МКЭ кажется целесообразным получить именно здесь, развивая метод Галёркина для краевой задачи (17.2)–(17.4).

Введем на отрезке $[a, b]$ равномерную сетку с шагом $h = \dfrac{b-a}{n+1}$, состоящую из n внутренних точек (узлов) $x_i = a + ih$ $(i = 1, 2, \ldots, n)$ и двух крайних — $x_0 = a$, $x_{n+1} = b$.

Будем искать приближенное решение $y_n(x)$ данной краевой задачи (17.2)–(17.4) в виде линейной комбинации простых однотипных функций $y_i(x)$, на роль которых возьмем так называемые ***финитные функции***, определяемые равенством

$$\varphi(t) = \begin{cases} 1 - |t|, & |t| \le 1, \\ 0, & |t| > 1. \end{cases} \qquad (17.66)$$

*) См. об этом далее в § 19.6.

Считая переменную x принадлежащей отрезку $[a, b]$ и полагая в (17.66) $t = \dfrac{x - x_i}{h}$, видим, что

$$|t| < 1 \quad \Leftrightarrow \quad x_i - h < x < x_i + h,$$

т.е. функция $\varphi(x)$ отлична от нуля лишь на интервале (x_{i-1}, x_{i+1}) с центром в точке x_i. Таким образом, разные узлы x_i рассматриваемой сетки определяют разные функции $\varphi_i(x)$, которые, очевидно, согласно (17.66), можно задать равенствами

$$\varphi_i = \begin{cases} 1 + \dfrac{x - x_i}{h} = \dfrac{x - x_{i-1}}{h}, & \text{если } x \in [x_{i-1}, x_i], \\ 1 - \dfrac{x - x_i}{h} = -\dfrac{x - x_{i+1}}{h}, & \text{если } x \in [x_i, x_{i+1}], \\ 0, & \text{если } x \notin [x_{i-1}, x_{i+1}]. \end{cases} \qquad (17.67)$$

График одной такой функции φ_i приведен на рис. 17.4.

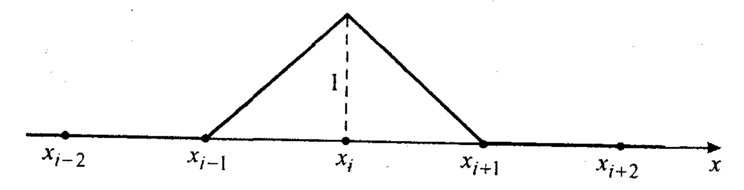


Рис. 17.4. График финитной функции φ_i

Сравнение рис. 17.4 с рис. 11.7 и формул (17.67) с формулами (11.62) показывает, что введенная здесь финитная функция, умноженная на постоянную $\dfrac{1}{h}$, совпадает с определенным в гл. 11 линейным B-сплайном.

Относительно совокупности финитных функций $\{\varphi_i\}$, задаваемых на отрезке $[a, b]$ формулами (17.67), известно, что *они линейно независимы* (более того, ортогональны в специальной энергетической норме, об этой норме см. далее в § 19.5, а также в [4, 47, 126]) *и образуют полную систему**) *в пространстве* $L_2[a, b]$. Это дает основание для их законного применения в качестве базисных функций метода Галёркина.

Для дальнейшей конкретизации описанного в общих чертах МКЭ нужно уточнить вид системы линейных алгебраических уравнений (17.59) относительно коэффициентов c_i, с которыми

*) Говоря о полноте, имеем в виду процесс $h \to 0$.

должны суммироваться выбранные базисные функции при получении приближенного решения, и из (17.60) вывести формулы коэффициентов этой системы. С этой целью продифференцируем (17.67), получив при этом

$$\varphi_i' = \begin{cases} \dfrac{1}{h} & \text{при} \quad x \in (x_{i-1}, x_i), \\ -\dfrac{1}{h} & \text{при} \quad x \in (x_i, x_{i+1}), \\ 0 & \text{при} \quad x \notin (x_{i-1}, x_{i+1}), \end{cases} \tag{17.68}$$

и изобразим параллельно всю систему определенных на отрезке $[a, b]$ функций φ_i и их производных φ_i' (рис. 17.5).

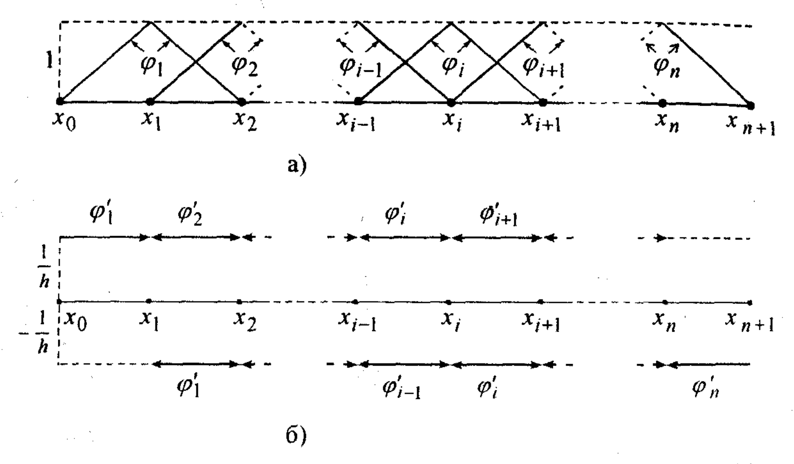


Рис. 17.5. Система финитных функций $\varphi_1, \varphi_2, \ldots, \varphi_n$ (а) и их производных (б)

Из совместного рассмотрения графиков φ_i и φ_i', наряду с уже отмеченным фактом, что i-я базисная функция и ее производная имеют ненулевые значения только на двух смежных элементарных промежутках (x_{i-1}, x_i) и (x_i, x_{i+1}), видим также, что на одном элементарном промежутке ненулевыми являются две «соседние» базисные функции и их производные: на (x_{i-1}, x_i) отличны от нуля φ_{i-1}, φ_i, φ_{i-1}', φ_i', на (x_i, x_{i+1}) — φ_i, φ_{i+1}, φ_i', φ_{i+1}', и т.д. Это позволяет, применив свойство аддитивности по промежутку интегрирования к интегралам формул (17.62), (17.63) метода Галёркина, упростить их вычисление, отбросив заведомо нулевые слагаемые, и увидеть определенную структуру матрицы линейной системы (17.59), что облегчает ее решение.

Итак, пусть ищется приближенное решение наиболее простой краевой задачи для дифференциального уравнения (17.2), а именно, с однородными краевыми условиями первого рода

$$y(a) = 0, \quad y(b) = 0. \tag{17.69}$$

Для получения ее приближенного решения в виде

$$y_n(x) = \sum_{i=1}^{n} c_i \varphi_i(x) \tag{17.70}$$

с кусочно-линейными базисными функциями $\varphi_i = \varphi_i(x)$, определенными в (17.67), для подсчета коэффициентов c_i, согласно методу Галёркина, нужно составить линейную алгебраическую систему (17.59). Ее правые части в таком случае суть

$$d_i = \int_a^b f(x)\varphi_i(x)dx = \sum_{k=0}^{n} \int_{x_k}^{x_{k+1}} f(x)\varphi_i(x)dx =$$

$$= \int_{x_{i-1}}^{x_i} f(x)\frac{x - x_{i-1}}{h}dx + \int_{x_i}^{x_{i+1}} f(x)\left(-\frac{x - x_{i+1}}{h}\right)dx =$$

$$= \frac{1}{h}\left[\int_{x_{i-1}}^{x_i} f(x)(x - x_{i-1})dx - \int_{x_i}^{x_{i+1}} f(x)(x - x_{i+1})dx\right]. \tag{17.71}$$

Так как при краевых условиях (17.69) используются n базисных функций с φ_1 по φ_n, и все они в точках a и b равны нулю (см. рис. 17.5), то формула (17.63) для вычисления элементов a_{ij} матрицы галёркинской СЛАУ (17.59) здесь имеет вид

$$a_{ij} = -\int_a^b \varphi_j'(x)\varphi_i'(x)dx + \int_a^b p(x)\varphi_j'(x)\varphi_i(x)dx + \int_a^b q(x)\varphi_j(x)\varphi_i(x)dx =$$

$$= \sum_{k=0}^{n} \int_{x_k}^{x_{k+1}} \left[-\varphi_j'(x)\varphi_i'(x) + p(x)\varphi_j'(x)\varphi_i(x) + q(x)\varphi_j(x)\varphi_i(x)\right]dx. \tag{17.72}$$

В силу отмеченного выше неравенства нулю на элементарном промежутке лишь соседних по индексу финитных функций и их производных, можно считать отличными от нуля фигурирующие в выражении a_{ij} произведения $\varphi_j'\varphi_i'$, $\varphi_j'\varphi_i$, $\varphi_j\varphi_i$, только в тех случаях, когда $i-1 \le j \le i+1$. А это означает, что

$$a_{ij} = 0 \quad \text{при} \quad |i - j| > 1, \tag{17.73}$$

т.е. матрица $\mathbf{A} = (a_{ij})$ системы (17.59) имеет ленточную структуру, точнее, является классической трехдиагональной матрицей. Это позволяет применять для ее решения высокоэффективный метод прогонки, как и в методе конечных разностей.

Конкретизируем формулы для вычисления ненулевых элементов матрицы $\mathbf{A}$. Полагая в (17.72) $j = i$, с помощью выражений (17.67), (17.68) получаем формулу для вычисления диагональных элементов:

$$a_{ii} = \int_{x_{i-1}}^{x_i} \left[-\frac{1}{h^2} + p(x)\frac{1}{h} \cdot \frac{x - x_{i-1}}{h} + q(x)\left(\frac{x - x_{i-1}}{h}\right)^2 \right] dx +$$

$$+ \int_{x_i}^{x_{i+1}} \left[-\left(-\frac{1}{h}\right)^2 + p(x)\frac{1}{h} \cdot \frac{x - x_{i+1}}{h} + q(x)\left(\frac{x - x_{i+1}}{h}\right)^2 \right] dx =$$

$$= -\frac{2}{h} + \frac{1}{h^2}\left[\int_{x_{i-1}}^{x_i} p(x)(x - x_{i-1})dx + \int_{x_{i-1}}^{x_i} q(x)(x - x_{i-1})^2 dx + \right.$$

$$\left. + \int_{x_i}^{x_{i+1}} p(x)(x - x_{i+1})dx + \int_{x_i}^{x_{i+1}} q(x)(x - x_{i+1})^2 dx \right]. \qquad (17.74)$$

При $j = i + 1$ из (17.72) находим выражения элементов правой побочной диагонали матрицы $\mathbf{A}$ системы (17.59):

$$a_{i,i+1} = \int_{x_i}^{x_{i+1}} \left[\frac{1}{h^2} + p(x)\frac{1}{h} \cdot \left(-\frac{x - x_{i+1}}{h}\right) + q(x)\left(-\frac{x - x_{i+1}}{h}\right) \cdot \frac{x - x_i}{h} \right] dx =$$

$$= \frac{1}{h} - \frac{1}{h^2}\left[\int_{x_i}^{x_{i+1}} p(x)(x - x_{i+1})dx + \int_{x_i}^{x_{i+1}} q(x)(x - x_i)(x - x_{i+1})dx \right], \qquad (17.75)$$

а при $j = i - 1$ — левой:

$$a_{i,i-1} = \int_{x_{i-1}}^{x_i} \left[\frac{1}{h^2} + p(x)\left(-\frac{1}{h}\right)\frac{x - x_{i-1}}{h} + q(x)\left(-\frac{x - x_i}{h}\right) \cdot \frac{x - x_{i-1}}{h} \right] dx =$$

$$= \frac{1}{h} - \frac{1}{h^2}\left[\int_{x_{i-1}}^{x_i} p(x)(x - x_{i-1})dx + \int_{x_{i-1}}^{x_i} q(x)(x - x_{i-1})(x - x_i)dx \right]. \qquad (17.76)$$

Таким образом, можно считать замысел МКЭ принципиально воплощенным, поскольку формулы (17.71), (17.73)–(17.76) полностью задают алгебраическую систему (17.59) для получения коэффициентов $c_1, \ldots, c_n$ приближенного решения (17.70) краевой задачи (17.2), (17.69). Сделаем лишь несколько замечаний.

Во-первых, находимая описанным способом функция $y_n(x)$ представляет собой *кусочно-линейную аппроксимацию* точного решения $y(x)$ краевой задачи (17.2), (17.69), а совокупность значений коэффициентов $c_1, \ldots, c_n$ играет роль *каркаса* приближенного решения $y_n(x)$ на сетке $x_1, \ldots, x_n$, что хорошо видно в результате подстановки $x = x_j$ в (17.70):

$$y_n(x_j) = \sum_{i=1}^{n} c_i \varphi_i(x_j) = c_j, \text{ в силу } \varphi_i(x_j) = \begin{cases} 1, & \text{если} \quad j = i, \\ 0, & \text{если} \quad j \neq i. \end{cases}$$

Во-вторых, при вычислении ненулевых элементов матрицы $\mathbf{A} = (a_{ij})$ линейной системы (17.59) следует учитывать, что в формулах (17.74)–(17.76) имеются одинаковые интегралы, что хотя бы незначительно сокращает затраты при численном интегрировании.

В-третьих, в данном случае галёркинскую систему алгебраических уравнений (17.59), учитывая (17.73), удобно представить как краевую задачу для трехточечного разностного уравнения второго порядка, т.е. в виде

$$\begin{cases} a_{11}c_1 + a_{12}c_2 = d_1, \\ a_{i,i-1}c_{i-1} + a_{ii}c_i + a_{i,i+1}c_{i+1} = d_i \quad (i = 2, \ldots, n-1), \\ a_{n,n-1}c_{n-1} + a_{nn}c_n = d_n, \end{cases} \qquad (17.77)$$

свидетельствующем о готовности системы к применению метода прогонки (2.23), (2.22).

В-четвертых, при неоднородных краевых условиях первого рода

$$y(a) = A, \quad y(b) = B \qquad (17.78)$$

можно воспользоваться описанным в § 17.4 приемом сведения задачи (17.2), (17.78) к задаче

$$L[u] = F(x),$$

где

$$F(x) = f(x) - p(x)v'(x) - q(x)v(x), \quad v(x) = A + \frac{B - A}{b - a}(x - a),$$

с однородными условиями

$$u(a) = 0, \quad u(b) = 0.$$

Найдя МКЭ ее приближенное решение

$$u_n(x) = \sum_{i=1}^{n} c_i \varphi_i(x),$$

получаем

$$y(x) \approx y_n(x) := u_n(x) + v(x).$$

Пример 17.3. Рассмотрим фигурирующее в примерах 17.1, 17.2 дифференциальное уравнение

$$y'' + x^2 y' - xy = \frac{6}{x^4} - \frac{3}{x}, \quad x \in [1, 2] \qquad (17.79)$$

при краевых условиях первого рода

$$y(1) = 1, \quad y(2) = 0.25. \qquad (17.80)$$

К этой задаче, имеющей то же точное решение $y(x) = \frac{1}{x^2}$, что и в предыдущей задаче (17.55) для этого уравнения, применим метод конечных элементов с двумя базисными функциями φ_1 и φ_2.

Сначала выполним преобразование данной задачи к задаче с однородными условиями

$$u'' + x^2 u' - xu = F(x), \quad x \in [1, 2], \quad u(1) = 0, \quad u(2) = 0, \qquad (17.81)$$

для чего делаем замену $y = u + v$, где

$$v = 1 + \frac{0.25 - 1}{2 - 1}(x - 1) = \frac{7}{4} - \frac{3}{4}x,$$

и, следовательно,

$$F(x) = \frac{6}{x^4} - \frac{3}{x} - x^2 v' + xv = \frac{6}{x^4} - \frac{3}{x} + \frac{7}{4}x.$$

Введя на отрезке $[1, 2]$ равномерную сетку

$$x_0 = 1, \quad x_1 = \frac{4}{3}, \quad x_2 = \frac{5}{3}, \quad x_3 = 2 \qquad (17.82)$$

с шагом $h = \frac{1}{3}$, записываем выражение приближенного решения задачи (17.81)

$$u_2(x) = c_1 \varphi_1(x) + c_2 \varphi_2(x),$$

где $\varphi_1(x)$ и $\varphi_2(x)$ — соответствующие сетке (17.82) функции-«крышки» (17.67). Для получения коэффициентов c_1, c_2 их линейной комбинации в

соответствии с (17.77) (при $n = 2$) составляем линейную алгебраическую систему*)

$$\begin{cases} a_{11}c_1 + a_{12}c_2 = d_1, \\ a_{21}c_1 + a_{22}c_2 = d_2. \end{cases} \qquad (17.83)$$

За числовыми данными этой системы обращаемся к формулам (17.71), (17.74)–(17.76), согласно которым имеем:

$$a_{11} = -6 + 9 \cdot \left[\int_1^{4/3} x^2 (x-1)dx + \int_1^{4/3} (-x)(x-1)^2 dx + \int_{4/3}^{5/3} x^2 \left(x - \frac{5}{3}\right) dx + \int_{4/3}^{5/3} (-x)\left(x - \frac{5}{3}\right)^2 dx \right] \approx -6.5926,$$

$$a_{22} = -6 + 9 \cdot \left[\int_{4/3}^{5/3} x^2 \left(x - \frac{4}{3}\right) dx + \int_{4/3}^{5/3} (-x)\left(x - \frac{4}{3}\right)^2 dx + \int_{5/3}^{2} x^2 (x-2)dx + \int_{5/3}^{2} (-x)(x-2)^2 dx \right] \approx -6.7407,$$

$$a_{12} = 3 - 9 \cdot \left[\int_{4/3}^{5/3} x^2 \left(x - \frac{5}{3}\right) dx + \int_{4/3}^{5/3} (-x)\left(x - \frac{4}{3}\right)\left(x - \frac{5}{3}\right) dx \right] \approx 3.9630,$$

$$a_{21} = 3 - 9 \cdot \left[\int_{4/3}^{5/3} x^2 \left(x - \frac{4}{3}\right) dx + \int_{4/3}^{5/3} (-x)\left(x - \frac{4}{3}\right)\left(x - \frac{5}{3}\right) dx \right] \approx 1.7037,$$

$$d_1 = 3\int_1^{4/3} \left(\frac{6}{x^4} - \frac{3}{x} + \frac{7}{4}x\right)(x-1)dx - 3\int_{4/3}^{5/3} \left(\frac{6}{x^4} - \frac{3}{x} + \frac{7}{4}x\right)\left(x - \frac{5}{3}\right) dx \approx 0.7256,$$

$$d_2 = 3\int_{4/3}^{5/3} \left(\frac{6}{x^4} - \frac{3}{x} + \frac{7}{4}x\right)\left(x - \frac{4}{3}\right) dx - 3\int_{5/3}^{2} \left(\frac{6}{x^4} - \frac{3}{x} + \frac{7}{4}x\right)(x-2)dx \approx 0.6446.$$

---

*) При таком малом числе базисных функций, разумеется, ни о какой ленточной структуре матрицы этой системы не может быть и речи; здесь даже нет ни одного «полного» трехточечного уравнения.

В результате подстановки этих чисел в систему (17.83) находим

$$c_1 \approx -0.1976, \quad c_2 \approx -0.1456.$$

Таким образом, приближенное решение $y_2(x)$ данной задачи (17.79), (17.80) есть

$$y_2(x) = v(x) + u_2(x) \approx 1.75 - 0.75x - 0.1976\varphi_1(x) - 0.1456\varphi_2(x),$$

где

$$\varphi_1(x) = \begin{cases} 3(x-1), & \text{при } x \in \left[1, \frac{4}{3}\right], \\ -3\left(x - \frac{5}{3}\right), & \text{при } x \in \left[\frac{4}{3}, \frac{5}{3}\right], \end{cases} \qquad (17.84)$$

$$\varphi_2(x) = \begin{cases} 3\left(x - \frac{4}{3}\right), & \text{при } x \in \left[\frac{4}{3}, \frac{5}{3}\right], \\ -3(x-2), & \text{при } x \in \left[\frac{5}{3}, 2\right]. \end{cases}$$

Последнее можно преобразовать в канонический вид кусочно-линейной на отрезке $[1, 2]$ функции

$$y_2(x) \approx \begin{cases} 2.3428 - 1.3428x, & \text{если } x \in \left[1, \frac{4}{3}\right], \\ 1.3444 - 0.5940x, & \text{если } x \in \left[\frac{4}{3}, \frac{5}{3}\right], \\ 0.8764 - 0.3132x, & \text{если } x \in \left[\frac{5}{3}, 2\right]. \end{cases} \qquad (17.85)$$

Чтобы получить представление о точности найденного приближенного решения $y_2(x)$, сравним его значения с точными во внутренних узлах сетки и в середине заданного промежутка, что отразим следующей таблицей:

x	$\frac{4}{3}$	$\frac{3}{2}$	$\frac{5}{3}$
$y_2(x)$	0.5524	0.4534	0.3544
$y(x)$	0.5625	0.4444	0.36

Изображения базисных функций φ_1, φ_2 (штрих-пунктирные линии), посредством которых методом конечных элементов получено кусочно-линейное приближение $y_2(x)$ к точному решению $y(x)$ данной в примере краевой задачи, и самих этих функций $y_2(x)$ и $y(x)$ (штриховая и сплошная линии соответственно) представлены на рис. 17.6.

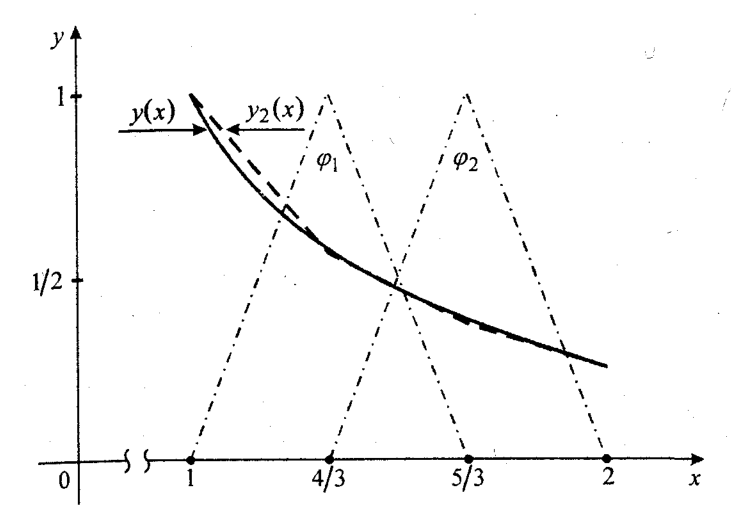


Рис. 17.6. *Графики точного решения задачи* (17.79)–(17.80), *приближения к нему МКЭ* (17.85) *и базисных функций* (17.84)

Увеличив количество узлов и базисных функций на единицу, приходим к СЛАУ

$$\begin{cases} -8.4167c_1 + 4.8333c_2 = 0.5992, \\ 2.9375c_1 - 8.5000c_2 + 5.1875c_3 = 0.4578, \\ 2.5417c_2 - 8.5833c_3 = 0.5020. \end{cases}$$

Найдя из нее значения c_1, c_2, c_3, получаем приближенное решение

$$y_3(x) \approx 1.75 - 0.75x - 0.1725\varphi_1(x) - 0.1806\varphi_2(x) - 0.1110\varphi_3(x)$$

с базисными функциями φ_1, φ_2, φ_3, легко конкретизируемыми посредством формулы (17.67) с учетом того, что узлами сетки в этом случае служат точки

$$x_0 = 1, \quad x_1 = 1.25, \quad x_2 = 1.5, \quad x_3 = 1.75, \quad x_4 = 2.$$

Приближение $y_3(x)$ уже существенно точнее построенного выше приближения $y_2(x)$: его каркас на данной сетке совпадает с каркасом точного решения $y(x)$ задачи (17.79), (17.80) во всех контролируемых здесь четырех десятичных знаках.

Замечание 17.7. Не касаясь особо вопросов, связанных со сходимостью и с точностью метода конечных элементов (доказано, что при определенных условиях можно считать $\|y(x) - y_n(x)\| = O(h^2)$), представим себе возможность его развития за счет использования B-сплайнов порядков, больше первого. Можно ожидать, что при таком повышении степени базисных функций будет повышаться точность приближенного решения с одновременным усложнением формул для подсчета коэффициентов линейной алгебраической системы, сохраняющей ленточную структуру, но с увеличивающейся шириной ленты из ненулевых элементов. Имеются также модификации МКЭ, приспособленные к эффективному решению краевых задач для дифференциальных уравнений с заведомо разрывными коэффициентами; первые представления об этом можно получить из учебного пособия [3]. Подробное описание МКЭ содержится, например, в книге [168].

Замечание 17.8. Декларируемая в начале параграфа равномерность сетки, обеспечивающая одинаковость ***конечных элементов***, коими в одномерном случае являются элементарные промежутки $[x_{i-1}, x_i]$, нужна была лишь для некоторого упрощения формул. Отказавшись от симметричности задания функции $\varphi(t)$ в (17.66), без большого труда можно распространить все рассуждения и выкладки на произвольные сетки (с переменным шагом $h_i := x_i - x_{i-1}$).

УПРАЖНЕНИЯ

17.1. Составьте алгоритм решения методом пристрелки (с точностью $\varepsilon > 0$ удовлетворения правому краевому условию) краевой задачи:

а) $x^2 y'' \ln(x) - xy' + y = 0, \quad x \in [1, e], \quad y(1) = 0, \quad y(e) = e - 2;$

б) $y'' = \sqrt{y'}, \quad x \in [0, 2], \quad y(0) = 0, \quad y(2) = \dfrac{2}{3}.$

17.2. К краевой задаче

$$y'' - y = x, \quad x \in [0, 1], \quad 2y(0) - y'(0) = 1, \quad y(1) = 2$$

примените: а) метод редукции;
б) метод дифференциальной прогонки.

17.3. Запишите формулы, определяющие для краевой задачи (17.2)–(17.4) метод п р а в о й дифференциальной прогонки (см. замечание 17.2). Примените их к решению задачи

$$y'' + \frac{1}{x} y' - \frac{3}{x^2} y = -\frac{3}{4\sqrt{x}}, \quad x \in [1, 4], \quad y(1) = 1, \quad y(4) - 2y'(4) = 2.$$

17.4. Методом конечных разностей второго порядка составьте алгебраическую систему уравнений относительно значений решения краевой задачи

$$y'' + 2y' - 3xy = \frac{2 - 8x}{x^3}, \quad x \in [1, 2], \quad y(1) = 1, \quad y(2) = 0.5 \qquad (17.86)$$

на сетке с шагом $h = 0.2$.

17.5. Дана краевая дифференциальная задача

$$y'' - y' \ln(x) - 2y = 1, \quad x \in [0.5, 1.5],$$

$$y(0.5) + y'(0.5) = 1, \quad y(1.5) - y'(1.5) = 0.$$

1) Аппроксимируйте ее разностной задачей с помощью МКР второго порядка на сетке с шагом $h = 0.125$. Подготовьте полученную СЛАУ к прогонке. Можно ли гарантировать устойчивость прогонки?
2) Примените противопотоковый метод на той же сетке.

17.6. Выполните конечноразностные аппроксимации краевой задачи

$$y'' + p(x)y' + q(x)y = f(x), \quad x \in [a, b], \quad y(a) = A, \quad y(b) = B$$

третьего и четвертого порядков точности. Какие можно отметить достоинства и недостатки построенных разностных схем? Опробуйте схему четвертого порядка на задаче (17.86).

17.7. Убедитесь в справедливости формул (17.51), (17.52) для подсчета параметров γ_i базисных функций (17.49), (17.50), опираясь на второе из однородных краевых условий (17.41).

17.8. Для краевой задачи (17.55) примера 17[illegible] (см. § 17.4), используя фигурирующие там функции φ_0 и φ_1, методом коллокации найдите квадратичное приближение $\hat{y}_1(x)$, взяв за основу другой узел коллокации $\hat{x}_1 = \dfrac{7}{4}$. Сравните значения $\hat{y}_1\left(\dfrac{3}{2}\right)$, $\hat{y}_1\left(\dfrac{7}{4}\right)$ и $\hat{y}_1(2)$ с соответствующими значениями точного решения $y(x)$ и найденного в примере 17.1 коллокацией в узле $x_1 = \dfrac{3}{2}$ приближенного решения $y_1(x)$.

17.9. Полагая $h = 0.25$, примените к краевой задаче (17.86):

а) метод коллокации со степенными базисными функциями;

б) метод Галёркина с теми же базисными функциями.

Сравните результаты в узлах сетки.

17.10. Выведите формулы для вычисления коэффициентов СЛАУ (17.58), к решению которой сводится применение метода Галёркина для дифференциального уравнения (17.65) с краевыми условиями первого рода (17.7) (см. замечание 17.5). На их основе получите соответствующие формулы метода конечных элементов для этой задачи.

17.11. Обобщите расчетные формулы метода конечных элементов так, чтобы они были пригодны для решения задачи (17.2), (17.78) в случае неравномерной сетки (см. замечание 17.8).

17.12. Дана краевая задача

$$(x+3)^2 y'' + (2x+6)y' + 0.25y = \sqrt{x+3}\,, \quad x \in [0,1],$$

$$y(0) = 0\,, \quad y(1) = 0.5\,.$$

А) Примените метод конечных элементов с одной, с двумя и с тремя базисными функциями на равномерной сетке. Сравните сеточные значения найденных приближений $y_1(x)$, $y_2(x)$, $y_3(x)$ и точного решения $y(x) = x/\sqrt{x+3}$.

Б) В случае единственного внутреннего узла (одной базисной функции) опробуйте результаты выполнения упр.17.11, принимая за этот узел поочередно точки $\frac{1}{3}$ и $\frac{2}{3}$ и сравнивая полученные приближения между собой и с $y_1(x)$ из п.А).

ГЛАВА 18 ЧИСЛЕННОЕ РЕШЕНИЕ ИНТЕГРАЛЬНЫХ УРАВНЕНИЙ

Даются некоторые общие понятия об интегральных уравнениях, в частности, о линейных уравнениях Фредгольма и Вольтерра. Описывается способ решения интегральных уравнений частного вида, известный как метод вырожденных ядер. Основное внимание уделяется численному решению уравнений методом конечных сумм, базирующемуся на использовании тех или иных квадратурных формул. Такой достаточно универсальный метод позволяет сводить данную задачу в функционально-интегральной постановке к решению линейной алгебраической системы относительно каркаса приближенного решения: с квадратной матрицей коэффициентов в случае уравнений Фредгольма и с треугольной — для уравнений Вольтерра. Изложение сопровождается численными примерами. Последний параграф главы посвящен квадратурно-итерационному методу вычисления каркаса резольвенты. Этот метод опирается на дискретизацию (на квадратурной основе) известной связи между резольвентой и ядром уравнения второго рода с последующим применением итерационного процесса Шульца.

18.1. НЕКОТОРЫЕ ОБЩИЕ СВЕДЕНИЯ ОБ ИНТЕГРАЛЬНЫХ УРАВНЕНИЯХ*)

Интегральным уравнением называют уравнение относительно неизвестной функции, содержащейся под знаком интеграла. С частными интегральными уравнениями мы уже встречались в главах 14, 15, преобразуя к интегральной форме задачу Коши для дифференциального уравнения (см. (14.3), (15.53)).

К интегральным уравнениям приводят многие задачи, возникающие и в самой математике, и в многочисленных ее приложениях. Исторически первой задачей, оформленной как интегральное уравнение

$$\int_0^z \frac{\varphi(\eta)}{\sqrt{z-\eta}}\, d\eta = f(z), \tag{18.1}$$

считается ***задача Абеля*** (1823 г.), заключающаяся в определении вида кривой $x = \varphi(z)$, по которой в вертикальной плоскости Oxz под действием силы тяжести скатывается материальная точка так, чтобы начав свое движение без начальной скорости в точке

*) Более подробно см. в источниках [41, 77, 95, 109, 110, 116, 130, 190 и др.]

кривой с аппликатой z, она достигала оси Ox за заданное время $T = f(z)$.

Другим примером может служить описание зависимостей между напряжениями σ и деформациями ε упруго-вязких материалов уравнениями

$$\varepsilon(t) = \frac{\sigma(t)}{E} + \frac{1}{E}\int_0^t K(t-\tau)\sigma(\tau)d\tau, \qquad (18.2)$$

$$\sigma(t) = E\varepsilon(t) - E\int_0^t T(t-\tau)\varepsilon(\tau)d\tau, \qquad (18.3)$$

где E — модуль упругости, $K(t-\tau)$ — функция влияния напряжения $\sigma(t)$ в момент τ на деформацию $\varepsilon(t)$ в момент t, $T(t-\tau)$ — аналогичная функция влияния деформации.

Основанием для составления уравнений обычно служат общие физические (в широком смысле) законы. Известные законы сохранения массы, импульса, энергии имеют интегральную формулировку и приводят к интегральным уравнениям. Интегральными уравнениями отражаются законы газовой динамики, электродинамики, экологии и т.д. Достоинством таких моделей служит то обстоятельство, что интегральные уравнения, в отличие от дифференциальных, не содержат производных искомой функции, а значит, не накладывают жестких ограничений на гладкость решения.

Достаточно общий вид интегральных уравнений содержится в записи

$$x(t) = \int_D K(t, s, x(s))ds + f(t), \qquad (18.4)$$

где D — некоторая область n-мерного пространства, x — неизвестная, а f — известная векторные функции; K — в общем случае нелинейная относительно x функция.

Ограничимся рассмотрением лишь одномерных уравнений, т.е. таких, в которых искомой неизвестной является скалярная функция одной переменной и интегрирование производится по отрезку. При этом будем иметь в виду только интеграл Римана, хотя часто для большей общности интегральные уравнения изучаются с привлечением интеграла Лебега. Более того, наше внимание не выйдет за рамки ***линейных интегральных уравнений***, т.е. таких уравнений (18.4), в которых подынтегральная функция $K(t, s, x(s))$ представима в виде $Q(t, s)x(s)$.

Опишем основные типы линейных интегральных уравнений. По тому, постоянны ли обе границы интегрирования или одна из них может быть переменной, линейные интегральные

уравнения подразделяются на уравнения Фредгольма и уравнения Вольтерра соответственно*). Более простыми (с более хорошими свойствами) и более широко применяемыми являются ***линейные интегральные уравнения второго рода***:

Фредгольма —

$$x(t) = \lambda\int_a^b Q(t, s)x(s)ds + f(t) \qquad (18.5)$$

и ***Вольтерра*** —

$$x(t) = \lambda\int_a^t Q(t, s)x(s)ds + f(t). \qquad (18.6)$$

Заданная функция $f(t)$ — свободный член — и неизвестная функция $x(t)$ — решение — в этих уравнениях зависят от переменной t, изменяющейся на отрезке $[a, b]$. Функция двух переменных $Q(t,s)$, называемая ***ядром*** интегрального уравнения, определяется на множестве точек квадрата $[a, b]\times[a, b]$ в случае интегрального уравнения Фредгольма (рис.18.1а) и треугольника $a \le s \le t \le b$ в случае уравнения Вольтерра (рис.18.1б).

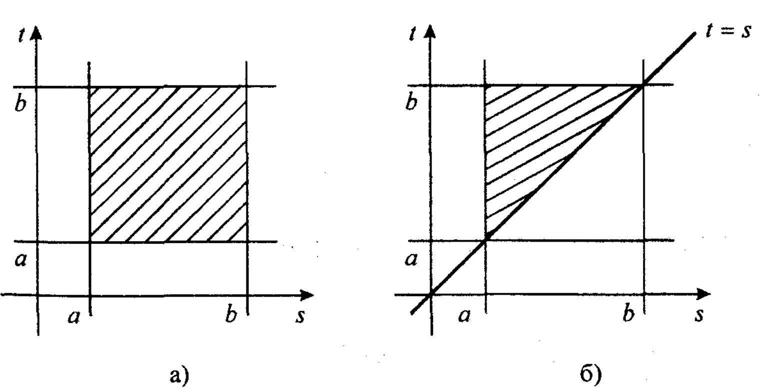


*Рис.*18.1. *Области задания ядер* $Q(t,s)$ *интегральных уравнений Фредгольма* (а) *и Вольтерра* (б)

*) Фредгольм Эрик Ивар (1866–1927) — шведский математик; Вольтерра Вито (1860–1940) — итальянский математик. Устоявшееся в течение многих лет русское написание «уравнение Вольтерра» все чаще сменяется написанием «уравнение Вольтерры». Отдадим дань традиции в ущерб правописанию.

Сразу заметим, что доопределив ядро $Q(t, s)$ уравнения Вольтерра нулем при $a \le s \le t \le b$ (в нештрихованной части квадрата $[a, b] \times [a, b]$ на рис.18.1б), уравнение Вольтерра можно посчитать уравнением Фредгольма и применять к нему известные для уравнений Фредгольма результаты. Однако при таком подходе теряется специфика уравнений Вольтерра, и возникают сложности, которых можно избежать при раздельном изучении этих уравнений*).

Введение в линейное интегральное уравнение числового параметра λ (который можно отнести и к ядру) придает уравнению более общий вид и позволяет установить теоремы существования решений при тех или иных значениях λ, в то время как при нейтральном значении $\lambda = 1$ решения может не оказаться. Подобно тому, как это делается в теориях линейных систем алгебраических или линейных дифференциальных уравнений, существование и единственность решений линейных неоднородных уравнений (18.5), (18.6) изучается посредством изучения соответствующих им однородных уравнений, т.е. уравнений, получающихся из (18.5), (18.6) при $f(t) \equiv 0$.

Более сложно обстоит дело с существованием решений, их единственностью и непрерывной зависимостью решений от правой части для ***интегральных уравнений первого рода***:

Фредгольма —

$$\int_a^b Q(t, s)x(s)ds = f(t) \tag{18.7}$$

и ***Вольтерра*** —

$$\int_a^t Q(t, s)x(s)ds = f(t). \tag{18.8}$$

Такие уравнения, характеризующиеся отсутствием отдельного слагаемого $x(t)$ (не связанного интегралом), имеют более ограниченную сферу применения и являются наиболее типичными представителями ***некорректных задач*** (см. § 1.7). Особенно это касается уравнений Фредгольма (18.7), в которых, кстати, иногда считают, что независимая переменная t изменяется не на промежутке интегрирования $[a, b]$, а на некотором другом отрезке $[c, d]$, где $c = a \pm \delta_1$, $d = b \pm \delta_2$, т.е. ядро $Q(t, s)$ задается на прямоугольнике $[c, d] \times [a, b]$.

*) Выше уже отмечалась связь между задачами Коши для дифференциальных уравнений и интегральными уравнениями с переменными верхними границами, относящимися к уравнениям Вольтерра. Аналогично, интегральные уравнения Фредгольма связывают с краевыми задачами для дифференциальных уравнений. Отсюда может проистекать понимание, например, того, насколько жестче стоит вопрос о разрешимости уравнения Фредгольма по сравнению с разрешимостью уравнения Вольтерра.

Уравнения первого и второго рода можно объединить общей записью

$$h(t)x(t) = \lambda \int_D Q(t, s)x(s)ds + f(t). \tag{18.9}$$

При $h(t) \equiv 0$ формула (18.9) определяет уравнение первого рода, при $h(t) \equiv 1$ — второго рода; если же $h(t)$ обращается в нуль в некоторых точках промежутка интегрирования D, то (18.9) есть ***уравнение третьего рода***, более редко встречающееся в приложениях и менее изученное.

Очень многие используемые на практике уравнения выделяются тем, что их ядра зависят только от разности $t - s$. Такие интегральные уравнения называют уравнениями с разностным ядром. К ним, очевидно, относятся приведенные в начале параграфа конкретные интегральные уравнения Вольтерра (18.1)–(18.3) (первое из них — уравнение первого рода, второе и третье — второго рода).

При непрерывных функциях $Q(t,s)$ и $f(t)$ достаточно просто устанавливается существование единственного непрерывного решения $x(t)$ уравнения Вольтерра второго рода (18.6) при любых значениях параметра λ; отсюда — несущественность его введения в эти уравнения. Для уравнений Фредгольма второго рода (18.5) при тех же требованиях непрерывности $Q(t,s)$ и $f(t)$ существование единственного непрерывного решения можно установить, например, при условии, что

$$|\lambda| < \frac{1}{C(b-a)}, \quad \text{где} \quad C := \max_{t, s \in [a, b]} |Q(t, s)|. \tag{18.10}$$

При снижении требований к гладкости возможных решений условие (18.10) ослабляется. Например, в случае функций, интегрируемых с квадратом, в роли достаточного условия вместо (18.10) часто фигурирует неравенство

$$|\lambda| < \frac{1}{\sqrt{\int_a^b \int_a^b Q^2(t, s)\,dt\,ds}}.$$

Для некоторых уравнений второго рода с ядрами определенной структуры имеются формулы (или совокупности формул), позволяющие найти точное решение $x(t)$.

Так, например, изучение уравнения Вольтерра (18.6) с $\lambda = 1$ и ***мультипликативным ядром***

$$Q(t, s) = p(t)\,q(s)$$

приводит к решению

$$x(t) = \int_a^t R(t, s)\, f(s)ds + f(t) \tag{18.11}$$

с функцией $R(t, s)$, подсчитываемой по формуле

$$R(t, s) = p(t)\,q(s)\,e^{\int_s^t p(\xi)\,q(\xi)d\xi} \;. \tag{18.12}$$

Решение уравнения Фредгольма (18.5) с ***вырожденным ядром***

$$Q(t, s) = \sum_{i=1}^{m} p_i(t)\,q_i(s) \tag{18.13}$$

имеет вид

$$x(t) = \lambda \sum_{i=1}^{m} z_i\; p_i(t) + f(t), \tag{18.14}$$

где числа z_i должны быть решениями системы линейных алгебраических уравнений

$$\begin{cases} z_1 - \lambda c_{11} z_1 - \lambda c_{12} z_2 - \ldots - \lambda c_{1m} z_m = d_1, \\ z_2 - \lambda c_{21} z_1 - \lambda c_{22} z_2 - \ldots - \lambda c_{2m} z_m = d_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ z_m - \lambda c_{m1} z_1 - \lambda c_{m2} z_2 - \ldots - \lambda c_{mm} z_m = d_m, \end{cases} \tag{18.15}$$

в которой

$$c_{ij} := \int_a^b q_i(s) p_j(s) ds\,, \quad d_i := \int_a^b q_i(s) f(s) ds\,. \tag{18.16}$$

Очевидно, существование единственного решения (18.14) уравнения (18.5) с ядром (18.13) напрямую связано с однозначной разрешимостью СЛАУ (18.15), т.е. с отличием от нуля определителя $D(\lambda)$ этой системы (называемого ***определителем Фредгольма***).

Пример 18.1. [96]. Рассмотрим уравнение

$$x(t) = \lambda \int_0^{\pi} \sin(t - s) x(s) ds + \cos(t)\,,$$

являющееся линейным интегральным уравнением Фредгольма второго рода. Его разностное ядро $\sin(t - s)$ по формуле синуса разности можно преобразовать к вырожденному:

$$\sin(t - s) = \sin t \cos s - \cos t \sin s\,.$$

Сопоставляя это с (18.13), видим, что если данное уравнение имеет решение, то его следует искать в соответствующей (18.14) форме:

$$x(t) = \lambda z_1 \sin t - \lambda z_2 \cos t + \cos t\,. \tag{18.17}$$

По формулам (18.16) вычисляем:

$$c_{11} := \int_0^{\pi} \cos s \sin s\, ds = 0\,, \qquad c_{12} := -\int_0^{\pi} \cos^2 s\, ds = -\frac{\pi}{2}\,,$$

$$c_{21} := \int_0^{\pi} \sin^2 s\, ds = \frac{\pi}{2}\,, \qquad c_{22} := -\int_0^{\pi} \sin s \cos s\, ds = 0\,,$$

$$d_1 := \int_0^{\pi} \cos^2 s\, ds = \frac{\pi}{2}\,, \qquad d_2 := \int_0^{\pi} \sin s \cos s\; ds = 0\,.$$

Составив систему вида (18.15)

$$\begin{cases} z_1 + \dfrac{\lambda\pi}{2} z_2 = \dfrac{\pi}{2}, \\ -\dfrac{\lambda\pi}{2} z_1 + z_2 = 0 \end{cases}$$

и вычислив определитель Фредгольма

$$D(\lambda) = \begin{vmatrix} 1 & \dfrac{\lambda\pi}{2} \\ -\dfrac{\lambda\pi}{2} & 1 \end{vmatrix} = 1 + \frac{\lambda^2\pi^2}{4}\,,$$

убеждаемся, что он отличен от нуля при любых вещественных значениях λ и, значит, решение (18.17) существует и единственно. Найдя еще определители

$$D_1(\lambda) = \frac{\pi}{2}\,, \qquad D_2(\lambda) = \frac{\lambda\pi^2}{4}\,,$$

по формулам Крамера записываем значения коэффициентов

$$z_1 = \frac{2\pi}{4 + \lambda^2\pi^2}\,, \qquad z_2 = \frac{\lambda\pi^2}{4 + \lambda^2\pi^2}\,,$$

подстановка которых в выражение (18.17) приводит к решению

$$x(t) = \frac{4\cos t + 2\lambda\pi \sin t}{4 + \lambda^2\pi^2} \qquad \forall \lambda \in \mathbf{R}.$$

Умение находить точное решение интегрального уравнения с вырожденным ядром порождает приближенный метод, в основе которого лежит замена одного уравнения другим, ядро которого вырожденно и в каком-то определеном смысле близко к данному. Такая замена ядра вырожденным может опираться на разные способы локальной аппроксимации функции двух переменных.

Например, уравнение

$$x(t) = \int_0^{0.5} t \sin(ts)\, x(s) ds + \cos 0.5t \qquad (18.18)$$

с определенной степенью точности (которую доступно оценить) можно заменить уравнением с вырожденным ядром

$$x(t) = \int_0^{0.5} t^2 s\, x(s) ds + \cos 0.5t, \qquad (18.19)$$

пользуясь тем, что по формуле Тейлора при малых ts имеет место приближенное равенство

$$\sin(ts) \approx ts + O((ts)^3). \qquad (18.20)$$

Кроме вышеупомянутого ***метода замены ядра на вырожденное***, имеется ряд других приближенно-аналитических методов решения интегральных уравнений. Из них отметим методы, тесно примыкающие к аналогичным методам решения дифференциальных уравнений. Например, сродни рассмотренному в § 14.1 методу Пикара решения задачи Коши является ***метод последовательных приближений*** для интегральных уравнений; аналогом метода Галеркина решения краевой задачи (см. § 7.5) служит ***метод моментов***. Описание этих и других приближенных методов можно найти в книгах [20, 41, 62, 81, 100, 116, 126, 162].

Наиболее универсальными и хорошо приспособленными для компьютерных вычислений являются численные методы решения интегральных уравнений. Их построение опирается на замену интеграла в интегральном уравнении конечной суммой на базе какой-либо квадратурной формулы, в результате чего задача сводится к алгебраической системе относительно дискретных значений (каркаса) искомого решения, соответствующих заданным или определяемых выбором квадратурной формулы значениям аргумента (сетке). Такие методы называются ***квадратурными методами*** или ***методами конечных сумм***. Они

чрезвычайно просты как в идее, так и в реализации, причем без каких-либо изменений их можно применить и к нелинейным интегральным уравнениям, имея лишь в виду, что в этом случае и системы конечномерных уравнений, к которым будет приводить дискретизация, также будут нелинейными.

Рассмотрению техники и особенностей применения метода конечных сумм к линейным интегральным уравнениям Фредгольма и Вольтерра посвящены следующие два параграфа.

18.2. КВАДРАТУРНЫЙ МЕТОД РЕШЕНИЯ ИНТЕГРАЛЬНЫХ УРАВНЕНИЙ ФРЕДГОЛЬМА

Пусть для вычисления определенного интеграла используется некая конкретная квадратурная формула

$$\int_a^b \varphi(s) ds \approx \sum_{j=1}^n A_j \varphi(s_j) \qquad (18.21)$$

с n узлами $s_j \in [a,b]$ и с соответствующими им весовыми коэффициентами A_j (см. гл.12).

Подставим правую часть приближенного равенства (18.21) с $\varphi(s) := Q(t, s)x(s)$ вместо интеграла в интегральное уравнение Фредгольма второго рода (18.5). В результате этого получаем

$$x(t) \approx \lambda \sum_{j=1}^n A_j Q(t, s_j) x(s_j) + f(t) \qquad (18.22)$$

— приближенное представление решения $x(t)$ уравнения (18.5) через n его значений $x(s_1)$, $x(s_2)$, ..., $x(s_n)$. Чтобы вычислить эти значения, станем рассматривать равенство (18.22) не при всех $t \in [a, b]$, а лишь на системе точек t_1, t_2, ..., t_n, совпадающих соответственно с узлами s_1, s_2, ..., s_n квадратурной формулы (18.21). Таким образом, приходим к n равенствам вида

$$x(t_i) \approx \lambda \sum_{j=1}^n A_j Q(t_i, s_j) x(s_j) + f(t_i), \qquad (18.23)$$

где $i = 1, 2, \ldots, n$.

Положим для краткости

$$Q_{ij} := Q(t_i, s_j), \quad f_i := f(t_i), \quad x_i :\approx x(t_i) \ (= x(s_i)). \quad (18.24)$$

Тогда (18.23) преобразуется к системе n линейных алгебраических уравнений с n неизвестными, которая выглядит следующим образом:

$$\begin{cases} (1-\lambda A_1 Q_{11})x_1 - \lambda A_2 Q_{12} x_2 - \ldots - \lambda A_n Q_{1n} x_n = f_1, \\ -\lambda A_1 Q_{21} x_1 + (1-\lambda A_2 Q_{22})x_2 - \ldots - \lambda A_n Q_{2n} x_n = f_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ -\lambda A_1 Q_{n1} x_1 - \lambda A_2 Q_{n2} x_2 - \ldots + (1-\lambda A_n Q_{nn})x_n = f_n. \end{cases} \qquad (18.25)$$

Стандартная запись линейной системы (18.25), с одной стороны, говорит о возможности ее решения тем или иным стандартным методом, если она совместна, с другой стороны, высвечивает роль параметра λ в однозначной разрешимости этой системы (и, как следствие, с некоторыми оговорками, интегрального уравнения (18.5)): при любых фиксированных A_j, Q_{ij} за счет уменьшения модуля параметра λ можно добиться диагонального преобладания в матрице системы, влекущего выполнимость условий сходимости метода простых итераций в форме Якоби § 3.2.

Найдя решение системы (18.25), получаем n чисел $x_1, x_2, \ldots, x_n$ — ***каркас приближенного решения*** интегрального уравнения на сетке $t = t_i$ $(i = 1, 2, \ldots, n)$. Если нам только и нужна эта таблица приближенных значений решения интегрального уравнения (18.5) в точках $t_1, t_2, \ldots, t_n$, то в такой постановке задача решена. Если же нужны значения решения в других точках $t \in [a, b]$ или требуется знать решение $x(t)$ в аналитическом виде, подставляем найденные значения $x_j \approx x(s_j)$ $(j = 1, 2, \ldots, n)$ в равенство (18.22) или аппроксимируем их, привлекая для этих целей, например, интерполяцию, если n мало, или сплайн-интерполяцию при больших n; в некоторых случаях здесь может оказаться полезным и метод наименьших квадратов.

Точность численного решения интегрального уравнения квадратурным методом зависит от нескольких взаимосвязанных факторов: применяемой квадратурной формулы, количества используемых узлов, свойств определяющих уравнение функций. Имеются неравенства, позволяющие, в принципе, оценить максимальное расхождение между точными и приближенными решениями при различных применяемых квадратурах (18.21) (см., например, [20, 41, 81, 100]); однако эти оценки малопригодны в рядовых ситуациях ввиду громоздкости, что, в конечном итоге, отражается на их точности. Более практичным, хотя и недостаточно строгим способом контроля точности подсчитываемого методом квадратур каркаса решения, может служить уже использовавшийся ранее (см. §§ 12.5, 13.2, 14.7) принцип Рунге сравнения значений приближенных решений в общих узлах двух

разных сеток, используемых в квадратурной формуле. Разумеется, при таком способе контроля точности приходится ограничиваться квадратурными формулами, опирающимися на равномерные сетки, или привлекать для сравнения результатов применения, например, формул Гаусса или Чебышева другие каркасы, подсчитывая значения приближенных решений в каких-то точках $t \in [a, b]$, отличных от узловых, по формуле (18.22).

При выборе конкретной квадратурной формулы (18.21) для использования в методе конечных сумм важно понимать, что лучшие результаты могут быть получены в том случае, когда ее алгебраический порядок точности хорошо согласуется со степенью гладкости ядра и свободного члена интегрального уравнения (предопределяющих гладкость искомого решения). Если ядро и свободный член не являются достаточно гладкими, то для дискретизации интегрального уравнения не следует применять высокоточных квадратур, а лучше ограничиться такими квадратурными формулами, как формулы трапеций или прямоугольников.

Пример 18.2. Дано уравнение

$$x(t) = \int_1^2 \frac{x(s)\,ds}{\sqrt{t+s^2}} + \sqrt{t+1} - \sqrt{t+4} + t. \qquad (18.26)$$

К этому интегральному уравнению Фредгольма второго рода применим метод конечных сумм на основе квадратурной формулы Гаусса с двумя узлами (учитываем, что на отрезке $[1,2]$ с гладкостью определяющих уравнение функций проблем нет). В соответствии с формулой (12.52) и табл.12.3 при $n = 2$ имеем

$$\int_1^2 \varphi(s)ds \approx \frac{1}{2}\left[\varphi\left(\frac{3-0.577350}{2}\right) + \varphi\left(\frac{3+0.577350}{2}\right)\right].$$

Используем это приближенное равенство в (18.26) с $\varphi(s) = \dfrac{x(s)}{\sqrt{t+s^2}}$. Приходим к приближенному представлению данного уравнения в виде

$$x(t) \approx \frac{x(1.211325)}{2\sqrt{t+1.211325^2}} + \frac{x(1.788675)}{2\sqrt{t+1.788675^2}} + \sqrt{t+1} - \sqrt{t+4} + t. \qquad (18.27)$$

Полагая здесь аргумент t равным узловым значениям $t_1 = 1.211325$ и $t_2 = 1.788675$ (к которым мы пришли от узлов ∓ 0.577350 в результате сопоставления стандартного промежутка интегрирования $[-1, 1]$ с данным промежутком $[1, 2]$), получаем линейную систему

$$\begin{cases} 0.694498x_1 - 0.238077x_2 = 0.415545, \\ -0.277095x_1 + 0.776125x_2 = 1.052641 \end{cases}$$

типа системы (18.25) относительно неизвестных $x_1 \approx x(t_1)$, $x_2 \approx x(t_2)$. Решив эту систему, находим каркас приближенного решения

$$x_1 \approx 1.2116, \quad x_2 \approx 1.7888.$$

Восполнение этого каркаса по формуле (18.27) приводит к аналитическому виду приближенного решения уравнения (18.26)

$$x(t) \approx \tilde{x}(t) := \frac{0.6058}{\sqrt{t+1.4673}} + \frac{0.8944}{\sqrt{t+3.1994}} + \sqrt{t+1} - \sqrt{t+4} + t.$$

Чтобы получить представление о точности полученного приближенного решения, сравним его с известным точным решением $x(t) = t$ в узлах сетки (т.е. сравним каркасы, определяемые выбранной квадратурной формулой), а также в точках, служащих концами и серединой промежутка интегрирования (каркасы, представляющие потенциальный интерес). Результаты сравнения отражены следующей таблицей:

$t\ (= x(t))$	1	1.211325	1.5	1.788675	2
$\tilde{x}(t)$	1.0003	1.2116	1.5002	1.7888	2.0001

В случае применения простейшей формулы Симпсона (12.19), опирающейся здесь на узлы $t_1 = 1$, $t_2 = 1.5$, $t_3 = 2$, приходим к представлению решения

$$x(t) \approx \frac{1}{6}\left[\frac{x(1)}{\sqrt{t+1}} + \frac{4x(1.5)}{\sqrt{t+2.25}} + \frac{x(2)}{\sqrt{t+4}}\right] + \sqrt{t+1} - \sqrt{t+4} + t$$

и соответственно к линейной алгебраической системе

$$\begin{cases} 0.88215x_1 - 0.36980x_2 - 0.07454x_3 = 0.17815, \\ -0.10541x_1 + 0.65574x_2 - 0.07107x_3 = 0.73593, \\ -0.09623x_1 - 0.32338x_2 + 0.93196x_3 = 1.28256. \end{cases}$$

Ее решение дает значения

$$x_1 \approx 0.9996, \quad x_2 \approx 1.4997, \quad x_3 \approx 1.9998,$$

также являющиеся достаточно хорошими приближениями к точным значениям $x(1) = 1$, $x(1.5) = 1.5$, $x(2) = 2$.

Замечание 18.1. При численном решении интегральных уравнений с разрывными ядрами или с бесконечными границами интегрирования следует вспомнить о наличии специально приспособленных для таких случаев квадратурных формул Гаусса–Кристоффеля, описанию которых посвящен § 12.7. Сделав правильный выбор квадратурной формулы из указанного семейства, все остальное можно выполнять по рассмотренной выше схеме.

Формально к интегральному уравнению Фредгольма первого рода (18.7) можно попытаться применить тот же метод конечных сумм, но результаты такого непосредственного перехода к квадратурам могут оказаться весьма неудовлетворительными из-за уже упоминавшейся в предыдущем параграфе некорректности этих уравнений. Поэтому прежде, чем осуществлять дискретизацию уравнения первого рода с помощью каких-либо квадратурных формул, производят его регуляризацию. Наиболее типичные предположения, в которых это делается, следующие.

Само уравнение рассматривается в записи

$$\int_a^b Q(t,s)x(s)ds = f(t), \quad t \in [c,d],$$

подчеркивающей тот факт, что промежуток изменения переменной t в общем случае может не совпадать с промежутком интегрирования (а это означает неквадратность матрицы коэффициентов СЛАУ, если сразу приступить к дискретизации этого уравнения). Ядро $Q(t,s)$ уравнения предполагается непрерывным в прямоугольнике $[c,d]\times[a,b]$, и считается, что вместо него известно близкое ему ядро $\tilde{Q}(t,s)$; их близость задается неравенством

$$\|\tilde{Q}(t,s) - Q(t,s)\| \le \xi.$$

Аналогично, вместо $f(t) \in L_2[c,d]$ считается известной функция $\tilde{f}(t) \approx f(t)$, причем

$$\|\tilde{f}(t) - f(t)\| \le \delta.$$

Относительно искомого решения $x(s)$ делается предположение, что при $s \in [a,b]$ оно должно иметь почти всюду производную $x'(s)$, интегрируемую с квадратом, и при этом $x'(a) = x'(b) = 0$.

В указанных предположениях схема получения каркаса регуляризированного решения такова.

Данное уравнение заменяется уравнением

$$\tilde{A}x := \int_a^b \tilde{Q}(t,s)x(s)ds = \tilde{f}(t), \quad t \in [c,d].$$

По этому интегральному оператору $\tilde{A}$ и функции $\tilde{f}$, согласно ***методу α-регуляризации Тихонова*** (см. § 1.7), строится ***функционал Тихонова***

$$\Phi_\alpha[x,\tilde{f}] := \int_a^b (\tilde{A}x - \tilde{f}(t))^2 dt + \alpha\,\Omega[x],$$

где ***стабилизирующий функционал*** часто берется в виде

$$\Omega[x] := \int_a^b \left(x^2(s) + q(x'(s))^2\right) ds,$$

содержащем дополнительный параметр $q \geq 0$; при значении $q = 0$ говорят о ***регуляризации нулевого порядка***, при $q > 0$ — ***первого порядка***. Из условия минимума функционала $\Phi_\alpha[x, \tilde{f}]$ составляется ***уравнение Тихонова*** (сравните с (1.31))

$$\alpha(x_\alpha(u) - qx''_\alpha(u)) + \int_a^b K(u, s)x_\alpha(s)ds = F(u), \quad u \in [a, b],$$

где

$$K(u, s) := \int_c^d \tilde{Q}(t, u)\tilde{Q}(t, s)dt \quad (= K(s, u)),$$

$$F(u) := \int_c^d \tilde{Q}(t, u)\tilde{f}(t)dt,$$

$x_\alpha(u)$ — искомое регуляризованное решение такое, что
$$x_\alpha(a) = 0, \ \ x'_\alpha(b) = 0.$$

Полученное интегро-дифференциальное уравнение (интегральное уравнение второго рода в случае $q = 0$) далее дискретизируется с помощью каких-либо квадратурных формул невысокого порядка точности; чаще всего здесь используется формула трапеций. После этого включаются непростые механизмы подбора оптимальных в каком-то смысле значений ***параметра регуляризации*** α, связанных с заданными значениями уровней погрешностей ξ и δ ядра и свободного члена. Подбор α завязан на многократном решении СЛАУ вида

$$(\alpha\mathbf{C} + \mathbf{G})\mathbf{x}_\alpha = \mathbf{F}$$

с симметричной положительно определенной матрицей $\alpha\mathbf{C} + \mathbf{G}$ (матрица $\mathbf{C}$ — ленточная, при $q = 0$ — скалярная)*).

Конкретные формулы, алгоритмы, программы и примеры, отражающие применение методов регуляризации к интегральным уравнениям, можно найти, например, в книге [41].

---

*) В некоторых случаях симметризацию выполняют искусственно, умножая уравнения, получаемые в процессе дискретизации, на постоянные множители, определяемые шагами дискретизации.

18.3. КВАДРАТУРНЫЙ МЕТОД РЕШЕНИЯ ИНТЕГРАЛЬНЫХ УРАВНЕНИЙ ВОЛЬТЕРРА

Выше отмечалось, что параметр λ в линейных интегральных уравнениях Вольтерра не несет такой нагрузки, как в случае уравнений Фредгольма. Поэтому положим в уравнении (18.6) $\lambda = 1$ и будем численно решать уравнение

$$x(t) = \int_a^t Q(t, s)x(s)ds + f(t), \quad t \in [a, b]. \qquad (18.28)$$

Учитывая, что это уравнение формально можно считать уравнением Фредгольма вида

$$x(t) = \int_a^b K(t, s)x(s)ds + f(t) \qquad (18.29)$$

с ядром

$$K(t, s) = \begin{cases} Q(t, s) & \text{при } a \leq s \leq t \leq b, \\ 0 & \text{при } a \leq t \leq s \leq b, \end{cases} \qquad (18.30)$$

для приближенного представления его решения

$$x(t) \approx \sum_{j=1}^{n} A_j Q(t, s_j)x(s_j) + f(t) \qquad (18.31)$$

на основе квадратурной формулы с узлами $s_j \in [a, b]$ и весами A_j можно воспользоваться результатами предыдущего параграфа. Согласно им, для получения каркаса $x_1, x_2, \ldots, x_n$ приближенного решения (18.31) нужно составить систему линейных алгебраических уравнений типа системы (18.25), которая применительно к уравнению (18.29), в силу (18.30), превращается в треугольную:

$$\begin{cases} (1 - A_1Q_{11})x_1 = f_1, \\ -A_1Q_{21}x_1 + (1 - A_2Q_{22})x_2 = f_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ -A_1Q_{n1}x_1 - A_2Q_{n2}x_2 - \ldots + (1 - A_nQ_{nn})x_n = f_n. \end{cases} \qquad (18.32)$$

Отсюда легко получаем последовательно один за другим иско-

мые значения $x_1, x_2, \dots, x_n$, полагая $i = 1, 2, \dots, n$ в формуле

$$x_i = \frac{f_i + \sum_{j=1}^{i-1} A_j Q_{ij} x_j}{1 - A_i Q_{ii}}. \qquad (18.33)$$

Применение такого формального подхода к численному решению уравнений Вольтерра имеет некоторые нюансы. Чтобы не разбираться с ними на этой стадии решения рассматриваемой задачи, проще заняться процессом дискретизации интегрального уравнения (18.28) с самого начала. При этом, учитывая, что точка b, ограничивающая промежуток определения решения $x(t)$, в уравнениях Вольтерра присутствует лишь номинально (как мы помним, эти уравнения родственны н а ч а л ь н ы м задачам для ОДУ), то при их дискретизации лучше изменить очередность фиксирования переменных t и s по сравнению с тем, как это делалось в случае уравнений Фредгольма.

Будем придавать аргументу t в уравнении (18.28) возрастающие значения $t_1, t_2, \dots, t_n \in [a, b]$. Получим n интегральных равенств

$$x(t_i) = \int_a^{t_i} Q(t_i, s) x(s) ds + f(t_i), \qquad (18.34)$$

промежутки интегрирования в которых при каждом i теперь постоянны и увеличиваются с ростом i на величину шага $h_i := t_i - t_{i-1}$. Заменив определенные интегралы в (18.34) конечными суммами

$$\sum_{j=1}^{i} A_j Q(t_i, s_j) x(s_j) \qquad (18.35)$$

с переменным числом узлов, равным i *), приходим к приближенным равенствам

$$x(t_i) \approx \sum_{j=1}^{i} A_j Q(t_i, s_j) x(s_j) + f(t_i). \qquad (18.36)$$

Если, как и прежде, при любых $i \in \{1, 2, \dots, n\}$

$$t_i = s_i \quad \text{и} \quad x_i :\approx x(t_i),$$

*) Что можно отразить введением второго индекса в обозначение весового коэффициента квадратурной формулы, т.е. всюду в пределах этого параграфа есть смысл использовать A_{ij} в роли A_j.

то (18.36) определяет линейную алгебраическую систему

$$x_i = \sum_{j=1}^{i} A_j Q_{ij} x_j + f_i, \qquad i = 1, 2, \dots, n, \qquad (18.37)$$

в точности совпадающую с формально выписанной выше системой (18.32).

Из последних рассуждений, в частности, вытекает понимание того, что *совокупность всех весовых коэффициентов* A_j *в к а ж д о й строке системы* (18.32) *должна быть полной для выбранной квадратурной формулы, т.е. при каждом* $i = 1, 2, \dots, n$ *должно иметь место равенство*

$$\sum_{j=1}^{i} A_j = t_i - a.$$

При конкретизации вида квадратурной формулы для замены определенных интегралов в равенствах (18.34) конечными суммами (18.35) *предпочтение здесь отдается формулам замкнутого типа, например, квадратурной формуле трапеций* (не обязательно с постоянным шагом h_i). В таком случае при $t_1 = a$ из (18.34) сразу следует

$$x_1 = f_1 \quad (= f(a)).$$

Далее, при $i = 2$ к интегралу $\int_a^{t_2} = \int_{t_1}^{t_2}$ в (18.34) применяется простейшая формула трапеций (12.17), в результате чего получаем второе уравнение системы (18.37):

$$x_2 = \frac{t_2 - a}{2}(Q_{21}x_1 + Q_{22}x_2) + f_2.$$

К следующему интегралу $\int_a^{t_3} = \int_{t_1}^{t_3}$ применяется уже составная формула трапеций (12.28), которая в более простом варианте с постоянным шагом $h = t_i - t_{i-1}$ приводит к уравнению

$$x_3 = \frac{h}{2}Q_{31}x_1 + hQ_{32}x_2 + \frac{h}{2}Q_{33}x_3 + f_3,$$

и так далее. Таким образом, в случае использования квадратурной формулы трапеций с постоянным шагом $h = t_i - t_{i-1}$ $(i = 2, 3, \dots, n)$ треугольная система для получения каркаса

$x_1, x_2, \ldots, x_n$ приближенного решения линейного интегрального уравнения Вольтерра второго рода (18.28) приобретает вид*)

$$\begin{cases} x_1 = f_1, \\ -\frac{h}{2}Q_{21}x_1 + \left(1-\frac{h}{2}Q_{22}\right)x_2 = f_2, \\ -\frac{h}{2}Q_{31}x_1 - hQ_{32}x_2 + \left(1-\frac{h}{2}Q_{33}\right)x_3 = f_3, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ -\frac{h}{2}Q_{n1}x_1 - hQ_{n2}x_2 - \ldots - hQ_{n,n-1}x_{n-1} + \\ \qquad\qquad + (1-\frac{h}{2}Q_{nn})x_n = f_n. \end{cases} \qquad (18.38)$$

Нет никакой необходимости собирать сразу все требуемые для построения каркаса решения уравнения в систему, целесообразнее поочередно записывать их и по ходу разрешать относительно соответствующего неизвестного или просто последовательно пользоваться формулой для вычислений неизвестных $x_2, x_3, \ldots$ типа формулы (18.33).

Пример 18.3. Дано интегральное уравнение Вольтерра второго рода

$$x(t) = \int_0^t t\cos^2(ts^3)x(s)ds + t^2 - \frac{1}{3}tg(t^4). \qquad (18.39)$$

Применим к нему квадратурную формулу трапеций с шагом $h = 0.1$, используя четыре узла:

$$t_1 = s_1 = 0, \quad t_2 = s_2 = 0.1, \quad t_3 = s_3 = 0.2, \quad t_4 = s_4 = 0.3.$$

В соответствии с этой сеткой, рис.18.2, и видом отвечающих данному случаю уравнений (18.38) имеем:

$$x_1 = f(0) = 0,$$

$$x_2 = \frac{0.01 - \frac{1}{3}\text{tg }0.0001}{1 - 0.005\cdot\cos^2 0.0001} \approx 0.010017,$$

*) В случае переменного шага ее вид незначительно сложнее.

$$x_3 \approx \frac{0.04 - \frac{1}{3}\text{tg}0.0016 + 0.02\cos^2 0.0002\cdot 0.010017}{1 - 0.01\cos^2 0.0016} \approx 0.040068,$$

$$x_4 \approx \frac{0.09 - \frac{1}{3}tg0.0081 + 0.03\cos^2 0.0003\cdot 0.010017}{1 - 0.015\cos^2 0.0081} +$$
$$+ \frac{0.03\cos^2 0.0024\cdot 0.040068}{1 - 0.015\cos^2 0.0081} \approx 0.090155.$$

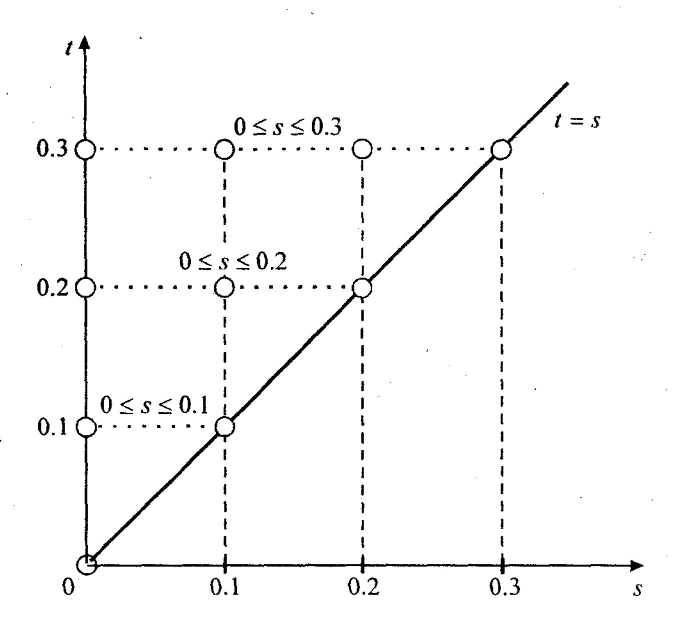


Рис. 18.2. Узлы и промежутки последовательного интегрирования при численном решении уравнения (18.39) методом трапеций с шагом $h = 0.1$

Полученные значения x_1, x_2, x_3, x_4 хорошо согласуются с соответствующими значениями $x(0), x(0.1), x(0.2), x(0.3)$ точного решения данного уравнения (18.39) $x(t) = t^2$. При этом замечаем, что с удалением от точки $a = 0$ точек t_i — фиксированных верхних границ интегрирования в выражениях вида (18.34) — точность уменьшается, что характерно и при численном нахождении решений задач Коши для ОДУ с постоянным шагом при удалении от начальной точки. Так как перемена очередности фикси-

рования переменных t и s в интегральном уравнении Вольтерра лишила нас аналитического представления приближенного решения через его каркас типа представления (18.22), применим лагранжеву интерполяцию. Составим по найденным числам x_1, x_2, x_3, x_4 — приближенным значениям решения $x(t)$ — таблицу конечных разностей:

i	t_i	x_i	Δx_i	$\Delta^2 x_i$	$\Delta^3 x_i$
1	0	0			
			0.010017		
2	0.1	0.010017		0.020034	
			0.030051		0.000002
3	0.2	0.040068		0.020036	
			0.050087		
4	0.3	0.090155			

Вторые разности в ней практически совпадают, поэтому следует ограничиться квадратичной интерполяцией. Взяв за основу первые три узла, по первой формуле Ньютона (1.25) получаем

$$x(t) \approx P_2(t) = x_1 + \frac{\Delta x_1}{h}(t - t_1) + \frac{\Delta^2 x_1}{2h^2}(t - t_1)(t - t_2) =$$

$$= 0 + \frac{0.010017}{0.1}t + \frac{0.020034}{2 \cdot 0.01}t(t - 0.1) = 1.0017t^2.$$

По трем следующим узлам (t_2, t_3, t_4) аналогично находим

$$x(t) \approx \widetilde{P}_2(t) = 1.0018t^2 - 0.00003t + 0.000002.$$

И то, и другое выражения могут служить неплохими аппроксимациями точного решения $x(t)$. Можно рассчитывать на лучшие результаты, если рассматривать сетку с более мелким шагом h.

Для решения уравнений Вольтерра первого рода (18.8) *удобнее применять квадратурные формулы открытого типа, например, формулу средних прямоугольников.* Фиксируя в таком уравнении переменную t равной значениям

$$t_1 = a + h, \quad t_2 = t_1 + h, \quad \ldots, \quad t_i = t_{i-1} + h, \tag{18.40}$$

получаем равенства

$$\int_a^{t_i} Q(t_i, s)x(s)ds = f(t_i), \quad i = 1, 2, \ldots, \tag{18.41}$$

из которых видно, что здесь нет необходимости считать узлы квадратур s_i совпадающими с узлами t_i; для формулы прямоугольников (средней точки) (12.8) берем их посередине элементарных промежутков интегрирования $[t_{i-1}, t_i]$ (рис 18.3).

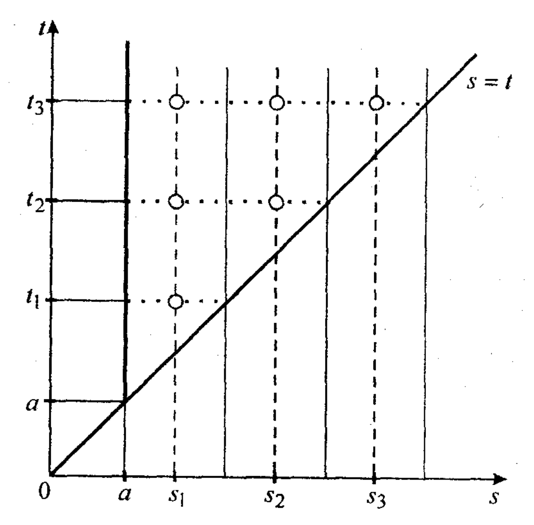


Рис. 18.3. Расположение узлов квадратурных формул прямоугольников при численном решении уравнений Вольтерра

При этом каркас приближенного решения определяется именно узлами s_i, т.е. полагаем

$$x_1 \approx x(s_1), \quad x_2 \approx x(s_2), \quad \ldots, \quad x_n \approx x(s_n). \tag{18.42}$$

При взятой за основу квадратурной формуле прямоугольников с учетом (18.42) равенства (18.41) приводят к следующим:

$$\begin{aligned} &hQ(t_1, s_1)x_1 = f(t_1), \\ &hQ(t_2, s_1)x_1 + hQ(t_2, s_2)x_2 = f(t_2), \\ &hQ(t_3, s_1)x_1 + hQ(t_3, s_2)x_2 + hQ(t_3, s_3)x_3 = f(t_3), \end{aligned} \tag{18.43}$$

и т.д. Напомним, что здесь узлы $t_1, t_2, \ldots$ сетки на оси Ot определены в (18.40), а

$$s_i := t_i - \frac{h}{2} \quad \forall \ i = 1, 2, \ldots .$$

Из равенств (18.43) последовательно находим:

$$x_1 = \frac{f(t_1)}{hQ(t_1, s_1)},$$

$$x_i = \frac{f(t_i) - h\sum_{j=1}^{i-1} Q(t_i, s_j)x_j}{hQ(t_i, s_i)} \quad (i = 2, 3, \ldots). \tag{18.44}$$

В случае применения в (18.41) квадратурных формул замкнутого типа при совпадающих системах узлов $\{s_i\}$ и $\{t_i\}$ возникает проблема вычисления значения $x_1 \approx x(t_1) = x(s_1) = x(a)$, которой не было в аналогичной ситуации с уравнениями второго рода. Действительно, это значение не может быть найдено непосредственно из равенства (18.41), при $i = 1$ теряющего смысл, а без него нельзя подсчитывать последующие значения x_2, x_3, ...

Чтобы вычислить x_1, продифференцируем рассматриваемое уравнение (18.8) по t и в полученном таким образом уравнении второго рода

$$Q(t, t)x(t) + \int_a^t Q_t'(t, s)x(s)ds = f'(t)$$

положим $t = a$. Имеем равенство

$$Q(a, a)x(a) = f'(a),$$

что равносильно $x(a) = f'(a)/Q(a, a)$, т.е. можно принять

$$x_1 = \frac{f'(a)}{Q_{11}}. \tag{18.45}$$

Далее при использовании, например, квадратурной формулы трапеций следует:

$$\frac{h}{2}Q_{21}x_1 + \frac{h}{2}Q_{22}x_2 = f_2 \quad \Rightarrow \quad x_2 = \frac{f_2 - \frac{h}{2}Q_{21}x_1}{\frac{h}{2}Q_{22}}, \tag{18.46}$$

$$\frac{h}{2}Q_{31}x_1 + hQ_{32}x_2 + \frac{h}{2}Q_{33}x_3 = f_3 \quad \Rightarrow$$

$$x_3 = \frac{f_3 - \frac{h}{2}Q_{31}x_1 - hQ_{32}x_2}{\frac{h}{2}Q_{33}}, \tag{18.47}$$

т.е. в общем случае при любом $j = 2, 3, \ldots$

$$x(s_j) \approx x_j = \frac{f_j - \frac{h}{2}Q_{j1}x_1 - h\sum_{k=2}^{j-1} Q_{jk}x_k}{\frac{h}{2}Q_{jj}} \tag{18.48}$$

(обозначения те же, что и в (18.24)).

Пример 18.4. Рассмотрим численное решение уравнения

$$\int_1^t (t^2 + s^2 + 1)\, x(s)ds = t^2 - \frac{1}{t} \tag{18.49}$$

на промежутке [1, 1.3] по формулам трапеций и прямоугольников, полагая шаг $h = 0.1$. (Точное решение $x(t) = \frac{1}{t^2}$ со значениями

$$x(1) = 1, \quad x(1.1) \approx 0.82645, \quad x(1.2) \approx 0.6944, \quad x(1.3) \approx 0.59172,$$

$$x(1.05) \approx 0.90703, \quad x(1.15) \approx 0.75614, \quad x(1.25) = 0.64).$$

Применим сначала формулу трапеций. Зафиксируем значения

$$t_1 = 1, \quad t_2 = 1.1, \quad t_3 = 1.2, \quad t_4 = 1.3$$

и в получающихся при этом из (18.49) равенствах

$$\int_1^{t_i} (t_i^2 + s^2 + 1)\, x(s)ds = t_i^2 - \frac{1}{t_i}, \qquad i = 1, 2, 3, 4$$

узлами s_j квадратур считаем те же точки 1, 1.1, 1.2, 1.3. Вычислив по

формуле (18.45) начальное значение решения

$$x(1)=x_1=\frac{f'(1)}{Q(1,1)}=\left.\frac{2t+\dfrac{1}{t^2}}{2t^2+1}\right|_{t=1}=1,$$

последующие его значения на сетке $\{s_j\}$ находим по формулам (18.46)–(18.48). Имеем:

$$x(1.1)\approx x_2=\frac{1.1^2-\dfrac{1}{1.1}-0.05\left(1.1^2+1^2+1\right)\cdot 1}{0.05\left(1.1^2+1.1^2+1\right)}\approx 0.8211,$$

$$x(1.2)\approx x_3=\frac{1.2^2-\dfrac{1}{1.2}-0.05\left(1.2^2+1^2+1\right)\cdot 1-0.1\cdot\left(1.2^2+1.1^2+1\right)\cdot 0.8211}{0.05\left(1.2^2+1.2^2+1\right)}\approx$$

$$\approx 0.6957,$$

$$x(1.3)\approx x_4=\frac{1.3^2-\dfrac{1}{1.3}-0.05\left(1.3^2+1^2+1\right)\cdot 1-0.1\left(1.3^2+1.1^2+1\right)\cdot 0.8211}{0.05\left(1.3^2+1.3^2+1\right)}-$$

$$-\frac{0.1\left(1.3^2+1.2^2+1\right)\cdot 0.6957}{0.05\left(1.3^2+1.3^2+1\right)}\approx 0.5877.$$

Теперь обратимся к формуле прямоугольников. Для данного случая в равенства (18.43) или в вытекающие из них формулы (18.44) при заданном $h=0.1$ нужно подставлять следующие значения:

$$Q(t_1,s_1)=Q(1.1,1.05)=3.3125,\qquad f(t_1)=f(1.1)\approx 0.30091;$$
$$Q(t_2,s_1)=Q(1.2,1.05)=3.5425,$$
$$Q(t_2,s_2)=Q(1.2,1.15)=3.7625,\qquad f(t_2)=f(1.2)\approx 0.60667;$$
$$Q(t_3,s_1)=Q(1.3,1.05)=3.7925,$$
$$Q(t_3,s_2)=Q(1.3,1.15)=4.0125,$$
$$Q(t_3,s_3)=Q(1.3,1.25)=4.2525,\qquad f(t_3)=f(1.3)\approx 0.92077.$$

Пользуясь ими, последовательно вычисляем:

$$x(1.05)\approx x_1=\frac{0.30091}{0.1\cdot 3.3125}\approx 0.90841,$$

$$x(1.15)\approx x_2=\frac{0.60667-0.1\cdot 3.5425\cdot 0.90841}{0.1\cdot 3.7625}\approx 0.75712,$$

$$x(1.25)\approx x_3=\frac{0.92077-0.1\cdot 3.7925\cdot 0.90841-0.1\cdot 4.0125\cdot 0.75712}{0.1\cdot 4.2525}\approx 0.64071.$$

Сравнение тех и других серий приближенных значений решения с точными значениями демонстрирует достаточно высокую эффективность применяемых квадратур (как и следовало ожидать, несколько лучшие результаты показывает квадратурная формула средней точки).

18.4. КВАДРАТУРНО-ИТЕРАЦИОННЫЙ МЕТОД ПОСТРОЕНИЯ РЕЗОЛЬВЕНТ

Предположим, что мы находимся в ситуации, когда требуется решать серию интегральных уравнений второго рода с одним и тем же ядром $Q(t,s)$ и разными свободными членами $f(t)$ *). Для определенности будем считать, что речь идет об интегральном уравнении Фредгольма (18.5).В таком случае, подобно тому, как это делается для линейных алгебраических систем $\mathbf{Ax}=\mathbf{b}$, где решения $\mathbf{x}^*$ при разных правых частях $\mathbf{b}$ вычисляются через предварительно найденную обратную к $\mathbf{A}$ матрицу $\mathbf{A}^{-1}$ по формуле $\mathbf{x}^*=\mathbf{A}^{-1}\mathbf{b}$, здесь также имеется возможность формально подсчитывать соответствующие разным $f(t)$ решения $x^*(t)$ уравнения (18.5) по формуле вида

$$x^*(t)=\lambda\int_a^b R(t,\tau;\lambda)f(\tau)\,d\tau+f(t). \qquad (18.50)$$

В этой формуле роль разрешающего оператора (аналога обратному) играет функция $R(t,s;\lambda)$, называемая ***резольвентой*** интегрального уравнения или, быть может, точнее, резольвентой семейства ядер $\lambda Q(t,s)$ уравнения (18.5) **) .При непрерывных ядрах $Q(t,s)$ и значениях параметра λ, удовлетворяющих, например, условию (18.10), несложно установить существование единственной непрерывной функции $R(t,s;\lambda)$, к которой абсо-

---

*) Оговоренная ситуация возникает, например, при решении **нелинейного** интегрального уравнения модифицированным методом Ньютона [80, 95].

) Иногда вместо термина *резольвента* используют термин ***разрешающее ядро* [109].

лютно и равномерно при $t, s \in [a, b]$ сходится ***ряд Неймана***[*]

$$\sum_{k=1}^{\infty} \lambda^{k-1} Q_k(t, s), \tag{18.51}$$

определяемый через ***итерированные ядра*** $Q_k(t, s)$ такие, что

$$Q_1(t, s) \equiv Q(t, s), \quad Q_k(t, s) = \int_a^b Q(t, \tau) Q_{k-1}(\tau, s)\, d\tau. \tag{18.52}$$

Последовательное вычисление итерированных ядер по рекуррентной формуле (18.52) и подсчет частичных сумм ряда (18.51) позволяют, в принципе, получить сколь угодно хорошее приближение к резольвенте $R(t, s; \lambda)$. Но цена такого приближения будет слишком высока (за редкими исключениями, когда такую процедуру можно провести аналитически, а не численно). Более продуктивные способы приближенного нахождения резольвенты опираются на простые интегральные соотношения между резольвентой и ядром. Практически в любом пособии по интегральным уравнениям можно найти вывод таких соотношений; их вид

$$R(t, s; \lambda) = Q(t, s) + \lambda \int_a^b Q(t, \tau) R(\tau, s; \lambda)\, d\tau \tag{18.53}$$

и

$$R(t, s; \lambda) = Q(t, s) + \lambda \int_a^b R(t, \tau, \lambda) Q(\tau, s)\, d\tau. \tag{18.54}$$

Тот факт, что резольвента, согласно формулам (18.53), (18.54), может рассматриваться как решение интегрального уравнения с тем же ядром $Q(t, s)$ (при фиксировании в ней одного из двух аргументов), позволяет подходить к ее численному построению с тех же позиций, которые были заложены в § 18.2 при приближенном решении уравнений Фредгольма второго рода (18.5).

Остановившись на какой-либо квадратурной формуле вида (18.21), по ее узлам на отрезке $[a, b]$ и, соответственно, на квадрате $[a, b] \times [a, b]$ строим сетку, узлами которой служат точки $(t_i; s_j)$, лежащие на пересечении линий $t = t_i$, $s = s_j$

[*] Нейман Карл Готфрид (1832–1925) — немецкий математик. Его основные труды относятся к дифференциальным уравнениям и алгебраическим функциям.

$(i, j = 1, 2, \ldots, n)$ в пределах данного квадрата[*]. Применение выбранной квадратурной формулы к интегральному соотношению (18.54) приводит к равенству

$$R(t, s; \lambda) = Q(t, s) + \lambda \sum_{k=1}^{n} A_k R(t, \tau_k; \lambda) Q(\tau_k, s) + r(t, s),$$

где $r(t, s)$ — погрешность квадратуры. Отбросив эту погрешность (малости которой можно добиться разными способами, например, увеличением числа узлов n), переходим к приближенному уравнению относительно точной резольвенты $R(t, s; \lambda)$, которое, в свою очередь, заменяем точным уравнением относительно приближенной резольвенты $\widetilde{R}(t, s; \lambda) \approx R(t, s; \lambda)$:

$$\widetilde{R}(t, s; \lambda) = Q(t, s) + \lambda \sum_{k=1}^{n} A_k \widetilde{R}(t, \tau_k; \lambda) Q(\tau_k, s). \tag{18.55}$$

Из функционального уравнения (18.55) фиксированием переменных $t = t_i$, $s = s_j$ $(i, j = 1, 2, \ldots, n)$ получаем n^2 дискретных уравнений

$$\widetilde{R}(t_i, s_j; \lambda) = Q(t_i, s_j) + \lambda \sum_{k=1}^{n} A_k \widetilde{R}(t_i, \tau_k; \lambda) Q(\tau_k, s_j) \tag{18.56}$$

относительно сеточных значений приближенной резольвенты, т.е. ее ***каркаса*** $\widetilde{R}_{ij}(\lambda) := \widetilde{R}(t_i, s_j; \lambda)$.

Введем $n \times n$-матрицы

$$\hat{\mathbf{R}}(\lambda) = \left(\widetilde{R}(t_i, s_j; \lambda)\right)_{i,j=1}^{n}, \quad \hat{\mathbf{Q}} := \left(Q(t_i, s_j)\right)_{i,j=1}^{n},$$

$$\mathbf{A} := diag(A_k)_{k=1}^{n}.$$

Учитывая совпадение значений t_i, s_j, τ_k при совпадении индексов i, j, k, совокупность уравнений (18.56) теперь можно переписать в виде одного матричного уравнения

$$\hat{\mathbf{R}}(\lambda) = \hat{\mathbf{Q}} + \lambda \hat{\mathbf{R}}(\lambda) \mathbf{A} \hat{\mathbf{Q}}, \tag{18.57}$$

в чем нетрудно убедиться, выполняя непосредственно (в элементах) фигурирующие в (18.57) матричные операции.

[*] Можно пойти и обратным путем: сначала ввести сетку, например, равномерную, а уже к ней подобрать подходящую квадратурную формулу.

Формально матричное уравнение (18.57) легко разрешается относительно искомого каркаса приближенной резольвенты:

$$\hat{\mathbf{R}}(\lambda) = \hat{\mathbf{Q}}\left(\mathbf{E} - \lambda \mathbf{A}\hat{\mathbf{Q}}\right)^{-1}. \qquad (18.58)$$

Согласно лемме Неймана, матрица $\left(\mathbf{E} - \lambda \mathbf{A}\hat{\mathbf{Q}}\right)^{-1}$ существует и представима матричным рядом

$$\left(\mathbf{E} - \lambda \mathbf{A}\hat{\mathbf{Q}}\right)^{-1} = \mathbf{E} + \lambda \mathbf{A}\hat{\mathbf{Q}} + \lambda^2 \left(\mathbf{A}\hat{\mathbf{Q}}\right)^2 + \lambda^3 \left(\mathbf{A}\hat{\mathbf{Q}}\right)^3 + \ldots \qquad (18.59)$$

в том и только в том случае, когда спектральный радиус матрицы $\lambda \mathbf{A}\hat{\mathbf{Q}}$ меньше единицы. Предположим, что это условие выполнено, т.е.

$$\rho\left(\lambda \mathbf{A}\hat{\mathbf{Q}}\right) < 1. \qquad (18.60)$$

Тогда искомый каркас приближенной резольвенты, в силу (18.58) и (18.59), также можно записать в виде суммы матричного ряда:

$$\hat{\mathbf{R}}(\lambda) = \hat{\mathbf{Q}} + \lambda \hat{\mathbf{Q}}\mathbf{A}\hat{\mathbf{Q}} + \lambda^2 \hat{\mathbf{Q}}\left(\mathbf{A}\hat{\mathbf{Q}}\right)^2 + \lambda^3 \hat{\mathbf{Q}}\left(\mathbf{A}\hat{\mathbf{Q}}\right)^3 + \ldots \qquad (18.61)$$

Вместо получения приближений к $\hat{\mathbf{R}}(\lambda)$ подсчетом частичных сумм ряда (18.61), который может сходиться весьма медленно, вернемся к представлению (18.58) и для обращения матрицы $\mathbf{E} - \lambda \mathbf{A}\hat{\mathbf{Q}}$ применим быстросходящийся итерационный процесс Шульца (3.34).

Чтобы не усложнять записи, примем за основу наиболее простой и распространенный процесс уточнения элементов обратной матрицы $\mathbf{B}^{-1}$ по формулам

$$\begin{cases} \mathbf{U}_{k+1} = \mathbf{U}_k + \mathbf{U}_k \psi_k, \\ \psi_k = \mathbf{E} - \mathbf{B}\mathbf{U}_k, \quad k = 0, 1, 2, \ldots. \end{cases} \qquad (18.62)$$

Известно, что необходимым и достаточным условием сходимости последовательности матриц $\mathbf{U}_k$ к матрице $\mathbf{B}^{-1}$ является условие

$$\rho(\psi_0) < 1, \qquad (18.63)$$

и если начальное приближение $\mathbf{U}_0$ выбрано так, что

$$\|\psi_0\| \le q < 1,$$

то справедливы оценки погрешности

$$\left\|\mathbf{B}^{-1} - \mathbf{U}_k\right\| \le \frac{\|\mathbf{U}_k \psi_k\|}{1 - \|\psi_k\|} \le \frac{\|\mathbf{U}_0\|}{1 - q} q^{2^k}, \qquad (18.64)$$

характеризующие (18.62) как метод второго порядка (см. § 3.6).

При обращении матрицы $\mathbf{E} - \lambda \mathbf{A}\mathbf{Q}$ в записи итерационного процесса (18.62) изменяется лишь вторая строка (для подсчета невязки), т.е. приближения $\mathbf{U}_k$ к $\left(\mathbf{E} - \lambda \mathbf{A}\hat{\mathbf{Q}}\right)^{-1}$ можно находить по формулам

$$\begin{cases} \mathbf{U}_{k+1} = \mathbf{U}_k + \mathbf{U}_k \psi_k, \quad k = 0, 1, 2, \ldots, \\ \psi_k = \mathbf{E} - \left(\mathbf{E} - \lambda \mathbf{A}\hat{\mathbf{Q}}\right)\mathbf{U}_k, \end{cases} \qquad (18.65)$$

если выполняется условие (18.63). Учитывая, что при условии (18.60) матрица $\left(\mathbf{E} - \lambda \mathbf{A}\hat{\mathbf{Q}}\right)^{-1}$ может быть разложена в ряд по формуле (18.59), возьмем в качестве начальной матрицы $\mathbf{U}_0$ в процессе (18.65) первый член этого разложения, т.е. положим

$$\mathbf{U}_0 := \mathbf{E}.$$

Тогда начальная невязка, «величина» которой определяет сходимость метода, будет

$$\psi_0 = \mathbf{E} - \left(\mathbf{E} - \lambda \mathbf{A}\hat{\mathbf{Q}}\right)\mathbf{E} = \lambda \mathbf{A}\hat{\mathbf{Q}},$$

и, следовательно, в таком случае необходимое и достаточное условие (18.63) сходимости итерационного процесса (18.65) совпадает с необходимым и достаточным условием (18.60) существования искомой обратной матрицы и ее представимости матричным рядом (18.59).

Таким образом, *если* $\rho\left(\lambda \mathbf{A}\hat{\mathbf{Q}}\right) < 1$, *то начатый с* $\mathbf{U}_0 = \mathbf{E}$ *итерационный процесс* (18.65) *сходится к матрице* $\left(\mathbf{E} - \lambda \mathbf{A}\hat{\mathbf{Q}}\right)^{-1}$, *а значит, и последовательность матриц* $\mathbf{R}_k := \hat{\mathbf{Q}}\mathbf{U}_k$, *в силу равенства* (18.58), *сходится к матрице* $\hat{\mathbf{R}}(\lambda)$ *— каркасу приближенной резольвенты* $\widetilde{R}(t, s; \lambda)$ *при всяком фиксированном* λ, *отвечающем условию* $\rho\left(\lambda \mathbf{A}\hat{\mathbf{Q}}\right) < 1$. *Если, более того,*

$$\left\|\lambda \mathbf{A}\hat{\mathbf{Q}}\right\| \le q < 1, \qquad (18.66)$$

то можно воспользоваться вытекающими из (18.64) *оценками* (*апостериорной и априорной*)

$$\left\|\hat{\mathbf{R}}(\lambda) - \mathbf{R}_k\right\| \le \left\|\hat{\mathbf{Q}}\right\| \frac{\|\mathbf{U}_k \psi_k\|}{1 - \|\psi_k\|} \le \frac{\left\|\hat{\mathbf{Q}}\right\|}{1 - q} q^{2^k} \qquad (18.67)$$

при любых мультипликативных нормах таких, что $\|\mathbf{E}\| = 1$.

Подводя итог, скажем, что при определенных условиях, на-

пример, за счет измельчения сетки, искомая резольвента $R(t, s; \lambda)$ уравнения Фредгольма второго рода с ядром $\lambda Q(t, s)$ при фиксированных значениях λ может быть сколь угодно хорошо представлена функцией $\widetilde{R}(t, s; \lambda)$, т.е. $n \times n$-матрицей $\hat{\mathbf{R}}(\lambda)$; последняя, в свою очередь, может быть сколь угодно хорошо приближена матрицами $\mathbf{R}_k := \hat{\mathbf{Q}}\mathbf{U}_k$ с помощью квадратично сходящегося итерационного процесса (18.65) вычисления матриц $\mathbf{U}_k$ (предварительно проверяется условие (18.66)).

Пример 18.5. Рассмотрим поведение описанного процесса построения приближений к каркасу резольвенты, взяв ядро

$$Q(t,s) = t^2 s - ts^2, \quad 0 \le t,\ s \le 1 \tag{18.68}$$

и параметр $\lambda = 4$.

Воспользуемся квадратурной формулой Гаусса с двумя узлами

$$\int_0^1 \varphi(x)\,dx \approx 0.5\varphi(0.211325) + 0.5\varphi(0.788675)$$

(см. формулу (5.52) при $a = 0$, $b = 1$, $n = 2$) и в соответствии с ней зададим на квадрате $[0,1] \times [0,1]$ 2×2-сетку узлов $(t_i; s_j)$ искомого каркаса так, как это показано на рис. 18.4, где для удобства сопоставления сеточных значений и элементов матриц ось Ot «перевернута».

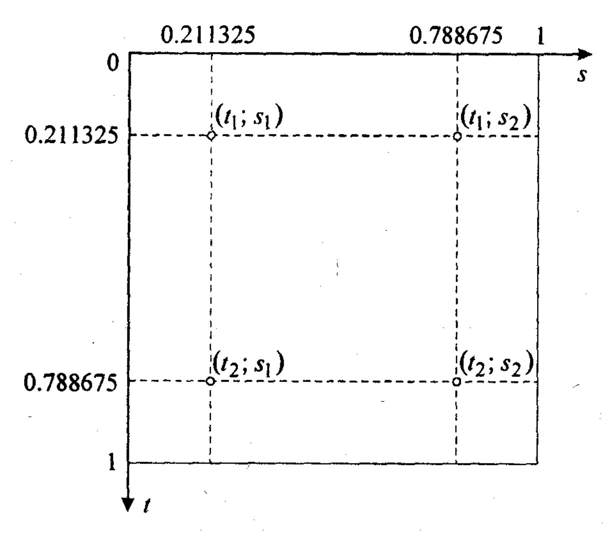


Рис.18.4. Сетка для подсчета каркаса резольвенты на основе квадратурной формулы Гаусса с двумя узлами

В таком случае имеем

$$\hat{\mathbf{Q}} \approx \begin{pmatrix} 0 & -0.0962 \\ 0.0962 & 0 \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}, \quad \mathbf{U}_0 := \mathbf{E} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

и, следовательно, при данном $\lambda = 4$

$$\psi_0 = \lambda \mathbf{A}\hat{\mathbf{Q}} = \begin{pmatrix} 0 & -0.1925 \\ 0.1925 & 0 \end{pmatrix}.$$

Убедившись в малости $\|\psi_0\|$, что означает выполнение требования (18.66) и, тем более, (18.63), далее продолжаем счет по формулам (18.65):

$$\mathbf{U}_1 = \begin{pmatrix} 1 & -0.1925 \\ 0.1925 & 1 \end{pmatrix}, \quad \psi_1 = \begin{pmatrix} -0.0370 & 0 \\ 0 & -0.0370 \end{pmatrix};$$

$$\mathbf{U}_2 = \begin{pmatrix} 0.9630 & -0.1853 \\ 0.1853 & 0.9630 \end{pmatrix}, \quad \psi_2 = \begin{pmatrix} 0.0014 & -0.0000 \\ 0.0000 & 0.0014 \end{pmatrix};$$

$$\mathbf{U}_3 = \begin{pmatrix} 0.9643 & -0.1856 \\ 0.1856 & 0.9643 \end{pmatrix}, \quad \psi_3 = \begin{pmatrix} 0.0000 & -0.0000 \\ 0.0000 & 0.0000 \end{pmatrix};$$

Ясно, что с используемым числом десятичных знаков дальнейшего уточнения значений элементов матриц $\mathbf{U}_k$, приближающих матрицу $\left(\mathbf{E} - \lambda \mathbf{A}\hat{\mathbf{Q}}\right)^{-1}$, происходить не будет. Значит, искомый каркас резольвенты ядра (18.68) с $\lambda = 4$ на заданной сетке может быть приближенно представлен матрицей

$$\mathbf{R}_3 = \hat{\mathbf{Q}}\mathbf{U}_3 = \begin{pmatrix} 0 & -0.0962 \\ 0.0962 & 0 \end{pmatrix}\begin{pmatrix} 0.9643 & -0.1856 \\ 0.1856 & 0.9643 \end{pmatrix} = \begin{pmatrix} -0.0179 & -0.0928 \\ 0.0928 & -0.0179 \end{pmatrix}.$$

Насколько полученная матрица $\mathbf{R}_3$ близка к истинному каркасу резольвенты на данной сетке, можно выяснить, зная для этого ядра точную резольвенту [В9]:

$$R(t,\ s;\ 4) = \frac{1}{8} ts(15t - 10ts - 6).$$

Подсчитав ее каркас (с четырьмя знаками после запятой)

$$\overline{\mathbf{R}} = \begin{pmatrix} -0.0183 & -0.0937 \\ 0.0867 & -0.0303 \end{pmatrix},$$

находим ошибку приближения $\mathbf{R}_3$:

$$\overline{\mathbf{R}} - \mathbf{R}_3 = \begin{pmatrix} -0.0004 & -0.0009 \\ -0.0059 & -0.0124 \end{pmatrix}.$$

Последняя матрица свидетельствует о том, что каркас резольвенты найден, грубо говоря, с точностью до сотых, и эта точность практически

достигается уже на первой итерации, т.е. можно принять

$$\overline{\mathbf{R}} \approx \mathbf{R}_1 = \hat{\mathbf{Q}}\mathbf{U}_1 = \begin{pmatrix} -0.0185 & 0.0962 \\ 0.0962 & -0.0185 \end{pmatrix}$$

с матрицей ошибок

$$\overline{\mathbf{R}} - \mathbf{R}_1 = \begin{pmatrix} 0.0002 & 0.0025 \\ -0.0095 & -0.0118 \end{pmatrix}.$$

Разобранный пример с известной резольвентой показывает, что при построении каркаса резольвенты квадратурно-итерационным методом точность итерационного обращения матрицы $\mathbf{E} - \lambda \hat{\mathbf{A}}\mathbf{Q}$ предопределяется точностью сведения интегральной задачи к матричной и легко регулируется указанием малости величин $\|\psi_k\|$ или $\|\mathbf{U}_k - \mathbf{U}_{k-1}\|$. Из-за проблем с реальным получением гарантированных оценок погрешностей дискретизации, здесь представляется целесообразным для контроля этих погрешностей квадратур использовать механизм сгущающихся равномерных сеток, общими узлами которых должны служить узлы заданной сетки искомого каркаса.

УПРАЖНЕНИЯ

18.1. По формулам (18.14)–(18.16) решите уравнение с вырожденным ядром (18.19). Сравните полученное таким способом приближенное решение уравнения (18.18) с его точным решением $x(t) \equiv 1$.

18.2. Запишите систему из n уравнений, к которой сводится применение метода конечных сумм на основе квадратурной формулы прямоугольников:

а) для *уравнения Гаммерштейна*

$$x(t) = \int_a^b Q(t, s)\varphi(s, x(s))\,ds, \quad t \in [a, b];$$

б) для *уравнения Урысона*

$$x(t) = \int_a^b K(t, s, x(s))ds, \quad t \in [a, b].$$

18.3. Составьте систему уравнений, к которой сводится применение метода конечных сумм на основе квадратурной формулы трапеций с пятью равностоящими узлами для решения *нелинейного уравнения Вольтерра*

$$x(t) = \int_a^t F(t, s, x(s))ds, \quad t \in [a, b].$$

18.4. Дано уравнение

$$x(t) = \int_1^2 \left(\frac{t}{s^2} - 1\right) x(s)ds = t^2 + \frac{t}{6} - \frac{7}{3}. \qquad (18.69)$$

Найдите его приближенное решение квадратурным методом с тремя узлами, пользуясь:

а) формулой трапеций;

б) формулой Гаусса.

В точках $t = 1$, $t = 1.5$ и $t = 2$ сравните полученные результаты с точными значениями решения $x(t)$, найдя его методом вырожденных ядер.

18.5. Дано уравнение Фредгольма первого рода

$$\int_0^1 \left(t^2 + s^3\right) x(s)ds = \frac{1}{2}t^2 + \frac{1}{8}, \quad t \in [0, 1].$$

1. Составьте для него уравнение Тихонова, полагая $\tilde{Q}(t, s) = t^2 + s^3$, $\tilde{f}(t) = \frac{1}{2}t^2 + \frac{1}{8}$.

2. Используя уравнение Тихонова, получите СЛАУ относительно значений каркаса α-регуляризированного решения на сетке 0, 0.5, 1, беря шаги дискретизации по всем независимым переменным одинаковыми $(= 0.5)$ и применяя квадратурную формулу трапеций:

а) в случае регуляризации нулевого порядка;

б) в случае регуляризации первого порядка (производные здесь аппроксимируйте разностными отношениями).

18.6. На уравнениях Вольтерра второго рода (18.39) и первого рода (18.49) примеров 18.3 и 18.4 соответственно, используя те же сетки, исследуйте применение в методе конечных сумм квадратурных формул левых и правых прямоугольников (12.6), (12.7). Сравните результаты между собой и с теми, которые были получены в указанных примерах с помощью других формул численного интегрирования.

18.7. Квадратурно-итерационным методом найдите приближенный каркас резольвенты для ядра уравнения (18.69) (считая $\lambda = 1$) на равномерной 3×3-сетке узлов квадрата $[1, 2] \times [1, 2]$, применяя:

а) квадратурную формулу трапеций;

б) квадратурную формулу Симпсона.

Предварительно оцените, сколько нужно сделать шагов процесса (18.65), чтобы вносимые неточным обращением матриц искажения каркаса резольвенты не превышали 0.001 в случае а и 0.0001 в случае б.

ГЛАВА 19 ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ С ЧАСТНЫМИ ПРОИЗВОДНЫМИ

Записываются и классифицируются основные уравнения математической физики и постановки задач для них. Напоминается один их немногих аналитических способов решения УМФ — метод Фурье, примененный здесь к задаче Дирихле для уравнения Лапласа в прямоугольнике. Описываются два подхода к построению полудискретного метода прямых (проекционный и конечноразностный); объектом для его демонстрации выбрана начально-граничная задача для уравнения теплопроводности. Далее рассматривается вариационная формулировка для операторного уравнения в гильбертовом пространстве, приводящая к энергетическому методу, и метод Ритца в развитие последнего. Показывается, к чему сводится реализация метода Ритца в случае решения им задачи Дирихле для уравнения Пуассона. Перечисляются наиболее важные моменты, отражающие суть метода конечных элементов, базирующегося на методе Ритца при фиксировании в нем в качестве координатных функций двумерных B-сплайнов.

19.1. ПРИМЕРЫ УРАВНЕНИЙ МАТЕМАТИЧЕСКОЙ ФИЗИКИ. КЛАССИФИКАЦИЯ УРАВНЕНИЙ С ЧАСТНЫМИ ПРОИЗВОДНЫМИ

При изучении большинства физических и иных процессов и явлений приходится сталкиваться с тем, что исследуемые свойства объекта описываются функциями не одной, а нескольких переменных величин. В таких случаях при составлении математических моделей изучаемых явлений вместо обыкновенных дифференциальных уравнений возникают уравнения с частными производными. Аргументам неизвестных функций таких уравнений зачастую придается смысл пространственных переменных и времени; тогда дифференциальные уравнения с частными производными по этим переменным, описывающие реальные физические модели (или идеальные физические явления), называются ***уравнениями математической физики***, а изучающая их наука — ***математической физикой*** [199].

Приведем несколько классических примеров уравнений математической физики. Искомой функцией в них выступает функция u, в разных задачах интерпретируемая по-разному, а аргументы x, y, z и t имеют смысл пространственных переменных и времени соответственно. Во многих уравнениях фигурирует сумма частных производных второго порядка, называемая

оператором Лапласа:

$$\Delta u := \operatorname{div}\operatorname{grad} u = \begin{cases} \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2}, & \text{если} \quad u = u(x, y), \\ \dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} + \dfrac{\partial^2 u}{\partial z^2}, & \text{если} \quad u = u(x, y, z). \end{cases} \tag{19.1}$$

1. ***Уравнение Лапласа*** (***уравнение потенциала***)

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0. \tag{19.2}$$

Это уравнение используется для математического описания плоских электростатических полей, магнитных полей постоянных токов, стационарных тепловых полей; применяют его и в задачах гидро- и аэродинамики. Аналогичные пространственные поля описываются ***трехмерным уравнением Лапласа***

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0. \tag{19.3}$$

Более коротко и единообразно двумерные и трехмерные уравнения Лапласа можно записать через оператор Лапласа (19.1):

$$\Delta u = 0. \tag{19.4}$$

Функции, удовлетворяющие уравнению Лапласа, называют гармоническими функциями [5, 53].

2. ***Уравнение Пуассона*** имеет вид

$$\Delta u = f, \tag{19.5}$$

где u и f одновременно являются (считаются) функциями либо двух, либо трех пространственных переменных. Область применения этого уравнения — задачи электростатики, электронной оптики, теории упругости и некоторые другие.

3. ***Уравнение теплопроводности*** (***Фурье***)

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2}, \tag{19.6}$$

где a — некоторая постоянная (в каждом конкретном случае своя) описывает диффузионные процессы, в частности, распространение тепла в тонком стержне. Очевидно, (19.6) — это одномерное уравнение теплопроводности. Для описания процесса распространения тепла в плоской пластинке или объемном теле

уравнение теплопроводности записывается через двумерный или, соответственно, трехмерный оператор Лапласа Δu:

$$\frac{\partial u}{\partial t} = a^2 \Delta u. \qquad (19.7)$$

Уравнение теплопроводности (19.6), а также уравнение (19.7), являются ***однородными уравнениями***. При наличии внутри стержня тепловых источников или поглотителей тепла обобщающее (19.6) уравнение теплопроводности будет неоднородным:

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2} + g(x, t).$$

4. ***Волновое уравнение (уравнение колебаний струны)*** — это уравнение вида

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}, \qquad (19.8)$$

описывающее свободные колебания, или

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2} + f(x, t),$$

если колебания совершаются под действием внешней силы, характеризуемой функцией $f(x, t)$. Двумерное волновое уравнение (***уравнение свободных колебаний однородной мембраны***) выглядит так:

$$\frac{\partial^2 u}{\partial t^2} = a^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right), \qquad (19.9)$$

что через двумерный оператор Лапласа можно записать короче в виде

$$\frac{\partial^2 u}{\partial t^2} = a^2 \Delta u.$$

5. ***Уравнение Гельмгольца***

$$\Delta u + cu = 0, \qquad (19.10)$$

где $c = const$, является математической моделью установившихся колебательных процессов. Имеется и неоднородное уравнение Гельмгольца — уравнение вида (19.10), дополненное правой частью.

6. ***Телеграфное уравнение***

$$\frac{\partial^2 u}{\partial t^2} + \frac{RC + LG}{LC} \frac{\partial u}{\partial t} + \frac{RG}{LC} u - \frac{1}{LC} \frac{\partial^2 u}{\partial x^2} = 0 \qquad (19.11)$$

описывает изменение потенциала u в линии электропередачи, где x — расстояние, t — время, L — коэффициент самоиндукции, C, R, G — соответственно емкость, сопротивление и характеристика потерь на единицу длины линии. При $R = 0$, $G = 0$ телеграфное уравнение (19.11) превращается в волновое уравнение (19.8).

7. ***Уравнения акустики*** — это система уравнений

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{1}{\rho_0} \dfrac{\partial p}{\partial x} = 0, \\ \dfrac{\partial p}{\partial t} + \rho_0 c_0^2 \dfrac{\partial u}{\partial x} = 0, \end{cases}$$

описывающая распространение плоских звуковых волн в покоящейся среде. Здесь: u — скорость возмущенной среды, p — давление в ней, ρ_0 — плотность, c_0 — постоянная, характеризующая сжимаемость среды.

8. ***Уравнение переноса***

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = f(x, t) \qquad (19.12)$$

служит простейшей одномерной моделью распространения частиц в веществе (где постоянную c можно интерпретировать как скорость переноса).

Приведенные здесь уравнения, разумеется, не исчерпывают все многообразие уравнений математической физики. Более подробные сведения об этих и других содержательных уравнениях в частных производных можно почерпнуть в специально посвященной им литературе и в некоторых пособиях по численным методам (см. [5, 9, 20, 21, 53, 78, 81, 103, 143, 144, 156, 160, 191] и др.). Главное, что приведенные выше уравнения являются наиболее простыми и, вместе с тем, характерными представителями множества уравнений математической физики. Это позволяет считать такие уравнения удобным и адекватным полигоном для изучения свойств уравнений в частных производных, корректных постановок задач с ними и построения методов их решения.

Анализируя форму вышеприведенных типичных уравнений, видим, что все они содержат частные производные искомой функции u не выше второго порядка, причем линейным образом (при этом, уравнения акустики, представляющие собой систему

линейных дифференциальных уравнений первого порядка, могут быть записаны с помощью одного линейного уравнения второго порядка). Следовательно, все те уравнения из перечисленных, в которых функция u зависит только от двух переменных, а именно, в случаях, когда $u = u(x, y)$ (см. 1, 2, 5), или $u = u(x, t)$ (см. 3, 4, 6–8), можно считать частными случаями уравнения

$$A\frac{\partial^2 u}{\partial x^2} + 2B\frac{\partial^2 u}{\partial x \partial y} + C\frac{\partial^2 u}{\partial y^2} + D\frac{\partial u}{\partial x} + E\frac{\partial u}{\partial y} + Gu = F. \qquad (19.13)$$

Запись (19.13) отражает общий вид ***линейного дифференциального уравнения с частными производными второго порядка*** относительно неизвестной функции $u(x, y)$. В наиболее общем случае его коэффициенты A, B, C, D, E, G и правая часть F есть некоторые функции независимых переменных x и y. Если $F \equiv 0$, то уравнение (19.13) называется ***однородным***; в противном случае — ***неоднородным***.

Классификацию конкретных уравнений в частных производных, принадлежащих семейству уравнений (19.13), производят по аналогии с классификацией кривых второго порядка в зависимости от знака ***дискриминанта*** $B^2 - AC$ (который, как доказано, не изменяется при различных естественных преобразованиях, т.е. знак дискриминанта является инвариантом таких преобразований). А именно, если $B^2 - AC < 0$, то уравнение (19.13) есть ***уравнение эллиптического типа***, если $B^2 - AC > 0$ — ***уравнение гиперболического типа***, если $B^2 - AC = 0$ — ***уравнение параболического типа***. При непостоянных коэффициентах A, B, C может оказаться, что дискриминант $B^2 - AC$ не имеет постоянного знака; в таком случае уравнение (19.13) относят к ***уравнениям смешанного типа***.

В случае постоянных коэффициентов при вторых производных в уравнении (19.13) с помощью простого преобразования переменных его приводят к виду, не содержащему смешанной производной, т.е. к той же форме (19.13), но с $B = 0$, и тогда определение типа уравнения весьма просто. Именно: *если в уравнении* (19.13) $B = 0$ *и постоянные коэффициенты при вторых производных имеют одинаковые знаки в одной части уравнения, то это уравнение — эллиптическое, если разные — гиперболическое; если вторая производная по одной из переменных отсутствует* (*т.е.* $B = 0$, $AC = 0$ *при* $A^2 + C^2 \neq 0$) — *параболическое.* Пользуясь этим правилом, легко классифицируем содержащиеся в этом параграфе конкретные уравнения математической физики, имеющие вторые частные производные относительно функции двух переменных: эллиптическими являются уравнения Лапласа, Пуассона, Гельмгольца; параболическим — уравнение теплопроводности; гиперболическими — волновое и телеграфное уравнения (к последнему типу также сводится система уравнений акустики). Уравнение переноса, хотя формально можно считать частным случаем уравнения (19.13), не содержит вторых частных производных и по дискриминанту не классифицируется.

Не составит труда записать общий вид линейного дифференциального уравнения второго порядка с частными производными искомой функции от трех и большего числа независимых переменных. Отнесение конкретных многомерных уравнений с частными производными к эллиптическому, параболическому или гиперболическому типу производят в зависимости от знака соответствующей квадратичной формы.

Для уравнений математической физики естественен и другой принцип классификации. В них, как уже отмечалось, аргумент t трактуется как время, а остальные независимые переменные играют роль пространственных координат. Поэтому в случае, когда уравнение с частными производными не содержит переменной t, оно называется ***стационарным***, при наличии в уравнении переменной t оно описывает процессы, развивающиеся во времени, и называется ***нестационарным*** или ***эволюционным уравнением*** (иногда понятие эволюционного уравнения трактуется несколько уже [117]).

19.2. ПОСТАНОВКИ ЗАДАЧ ДЛЯ УРАВНЕНИЙ МАТЕМАТИЧЕСКОЙ ФИЗИКИ

Каждое уравнение с частными производными, как и обыкновенное дифференциальное уравнение, имеет бесчисленное множество решений. При этом, если, например, линейное однородное обыкновенное дифференциальное уравнение второго порядка обладает базисом из двух линейно независимых функций-решений, через которые можно выразить любое его решение, введя произвольные постоянные, то для линейных однородных уравнений в частных производных второго порядка, каковыми являются уравнения (19.2), (19.3), (19.6), (19.11), такого конечного базиса не существует. Поэтому не только получение, но и формальная запись общего решения даже для простейших уравнений в частных производных зачастую вызывает большие затруднения. Однако постановщикам реальных задач, как правило, общее решение и не нужно. Интерес для них представляют те решения, которые обусловлены соответствующими уравнению данными, описывающими явление в целом. Ставя формальную задачу для конкретного уравнения в частных производных, следует позаботиться о том, чтобы добавляемые к уравнению из тех или иных соображений условия выделяли из общего решения единственное частное решение, чтобы это частное решение на

самом деле существовало в заданном функциональном пространстве и чтобы оно мало изменялось при малых изменениях условий. Эти три требования — разрешимости, однозначности и непрерывной зависимости от исходных данных (иначе, устойчивости) — в совокупности характеризуют ***корректность*** постановок задач математической физики (см. § 1.7).

Рассмотрим несколько примеров постановок задач для простейших из приведенных в предыдущем параграфе уравнений математической физики. Но прежде заметим, что для уравнений с частными производными так же, как и для обыкновенных дифференциальных уравнений, ставятся начальные и краевые задачи, хотя разделение это весьма условно. Если одна из независимых переменных играет роль времени, то условия, относящиеся к начальному моменту времени $(t = t_0)$, называются ***начальными условиями***, и соответствующая задача носит название ***начальной задачи*** (или ***задачи Коши****)). Условия, которые задаются при различных значениях пространственных переменных (обычно на границе области изменения этих переменных), называются ***граничными*** (или ***краевыми***) ***условиями***. Как правило, для уравнений эллиптического типа, описывающих стационарные процессы, задаются ***граничные условия***, т.е. ставятся ***граничные задачи***, а для уравнений параболического и гиперболического типов, моделирующих эволюционные процессы и явления, для определенности нужно одновременно задавать условия, начальные по времени и граничные по пространственным переменным, что приводит к ***смешанным задачам***.

Пусть линейное дифференциальное уравнение с частными производными второго прядка (19.13) является уравнением эллиптического типа (т.е. $B^2 - AC < 0$), и пусть $\Omega \subset \mathbf{R}_2$ — заданная двумерная область с границей $\mathbf{\Gamma}$, в замыкании которой определено уравнение и в которой ищется решение $u = u(x, y)$. Если наряду с уравнением (19.13) эллиптического типа выставляется условие

$$u\big|_{(x;\, y)\in\mathbf{\Gamma}} = \varphi(x, y), \qquad (19.14)$$

где $\varphi(x, y)$ — заданная во всех точках $(x; y)$ границы $\mathbf{\Gamma}$ непрерывная функция, то совокупность уравнения (19.13) и граничного условия (19.14) определяет ***первую краевую задачу***. Применительно к уравнениям Лапласа (19.2) и Пуассона (19.5) ее называют ***задачей Дирихле***. Если эллиптическое уравнение (19.13) рассматривается совместно с условием

$$\left.\frac{\partial u}{\partial \mathbf{n}}\right|_{(x;\, y)\in\mathbf{\Gamma}} = \varphi(x, y), \qquad (19.15)$$

где $\dfrac{\partial u}{\partial \mathbf{n}}$ — производная искомой функции $u(x, y)$ по направлению $\mathbf{n}$ внешней нормали, определяемой соответствующей точкой $(x; y)$ границы $\mathbf{\Gamma}$ заданной области Ω, то в таком случае имеем ***вторую краевую задачу***. В частности, для уравнений Лапласа и Пуассона вторую краевую задачу, т.е. сочетания (19.2) с (19.15) и (19.5) с (19.15), называют ***задачей Неймана***. Рассматривается и ***третья краевая задача*** для эллиптических уравнений, обобщающая первые две. Она определяется заданием дополнительного к уравнению (19.13) (при $B^2 - AC < 0$) условия

$$\left[\alpha_0 u + \alpha_1 \frac{\partial u}{\partial \mathbf{n}}\right]_{(x;\, y)\in\mathbf{\Gamma}} = \varphi(x, y),$$

где $|\alpha_0| + |\alpha_1| \neq 0$.

Примером постановки ***начально-граничной задачи*** *для гиперболического уравнения* может служить математическое описание процесса свободных колебаний однородной струны, закрепленной в точках $x = 0$ и $x = l$, которой в момент времени $t = 0$ сообщили начальное смещение $\varphi(x)$ и скорость $\psi(x)$. Процесс таких колебаний определяется уравнением (19.8) (где постоянная a^2 связана с весом и натяжением струны), сопровождаемым начальными

$$\left.\begin{aligned} &u(x, 0) = \varphi(x), \\ &\frac{\partial u(x, 0)}{\partial t} = \psi(x) \end{aligned}\right\}, \quad x \in [0, l] \qquad (19.16)$$

и граничными

$$\left.\begin{aligned} &u(0, t) = 0, \\ &u(l, t) = 0 \end{aligned}\right\}, \quad t \in [0, +\infty) \qquad (19.17)$$

условиями. При определенных требованиях к функциям $\varphi(x)$ и $\psi(x)$ решение $u(x, t)$ этой задачи позволяет в любой момент времени $t > 0$ для любой точки $x \in (0, l)$ струны однозначно указать величину u отклонения от положения равновесия ($u = 0$).

*) Здесь имеются определенные терминологические тонкости. Как отмечается в [53, с.91], часто в литературе начальной задаче для уравнения теплопроводности приписывается название задачи Коши, хотя таковой она не является.

Наконец, наиболее типичной задачей *для уравнения параболического типа* является следующая ***начально-граничная задача***:

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2}, \qquad (19.18)$$

$$u(x, 0) = \varphi(x) \quad \text{при} \quad x \in [0, l], \qquad (19.19)$$

$$u(0, t) = \alpha(t), \quad u(l, t) = \beta(t) \quad \text{при } t \geq 0. \qquad (19.20)$$

Физически она может интерпретироваться как задача определения температуры $u(x,t)$ в любой точке x тонкого однородного стержня длиной l в произвольный момент времени t, если известно распределение температуры в стержне в начальный момент времени $t = 0$ (см. начальное условие (19.19)) и известна температура на концах стержня $x = 0$ и $x = l$ в любой момент времени t (см. граничные условия (19.20)). Коэффициент a^2 в уравнении (19.18) связан с теплофизическими характеристиками материала стержня (зачастую уравнение вида (19.18) рассматривается без коэффициента a^2, что мотивируется превращением этого коэффициента в единицу при преобразовании уравнения к новой переменной $\Theta = a^2 t$).

Приведенные здесь задачи отражают основные черты множества постановок задач математической физики и далее будут служить объектами изучения с позиций численного анализа.

19.3. МЕТОД РАЗДЕЛЕНИЯ ПЕРЕМЕННЫХ

Одним из наиболее распространенных способов аналитического решения уравнений в частных производных является ***метод разделения переменных***, называемый также ***методом Фурье***. Этот метод впервые был предложен Ж.Даламбером в середине XVIII века для решения волнового уравнения (19.8) с начальными и краевыми условиями (19.16), (19.17), а в начале XIX века был развит Ж.Фурье и обоснован М.В.Остроградским и П.Дирихле [53, 199]. В любом учебном пособии, посвященном теории уравнений математической физики, можно найти описание метода Фурье разной степени полноты и строгости изложения (см., например, [5, 53]). Здесь мы покажем лишь идею этого метода, причем не на волновом уравнении, где он, возможно, смотрится более выигрышно, поскольку стали уже привычными ассоциации между колебательными процессами и тригонометрическими рядами Фурье, а на задаче Дирихле для уравнения Лапласа в прямоугольной области [81].

Пусть на плоскости Oxy задан прямоугольник

$\Omega := [0, a] \times [-b, b]$ (рис. 19.1). Граница $\mathbf{\Gamma}$ этой области $\mathbf{\Omega}$ образована четырьмя отрезками прямых

$$x = 0, \quad x = a, \quad y = -b \quad \text{и} \quad y = b,$$

на которых определены непрерывные функции

$$u = \psi_1(y), \quad u = \psi_2(y), \quad u = \varphi_1(x), \quad u = \varphi_2(x)$$

соответственно.

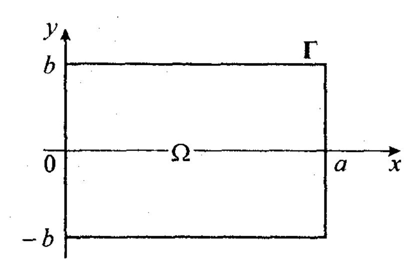


Рис. 19.1. *Область определения задачи Дирихле* (19.21)–(19.22)

Ставим задачу: найти функцию $u = u(x, y)$ такую, чтобы она удовлетворяла уравнению Лапласа

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad \text{при} \quad (x, y) \in \Omega \qquad (19.21)$$

и чтобы

$$u = \begin{cases} \psi_1(y) & \text{при} \quad x = 0, \\ \psi_2(y) & \text{при} \quad x = a, \\ \varphi_1(x) & \text{при} \quad y = -b, \\ \varphi_2(x) & \text{при} \quad y = b, \end{cases} \qquad (19.22)$$

где функции ψ_i, φ_j предполагаются согласованными в общих вершинах прямоугольника.

Учитывая линейность оператора Лапласа Δu, будем искать решение данной задачи в виде суммы двух функций: u_1 и u_2. Тогда, если каждая из них будет удовлетворять в Ω уравнению Лапласа (19.21) и, кроме того, по отдельности подчиняться условиям

$$u_1 = \begin{cases} \varphi_1(x) & \text{при} \quad y = -b, \\ \varphi_2(x) & \text{при} \quad y = b, \\ 0 & \text{при} \quad x = 0 \text{ и } x = a, \end{cases} \qquad (19.23)$$

и

$$u_2 = \begin{cases} \psi_1(y) & \text{при} \quad x = 0, \\ \psi_2(y) & \text{при} \quad x = a, \\ 0 & \text{при} \quad y = \pm b, \end{cases} \tag{19.24}$$

то функция $u = u_1 + u_2$ будет решением поставленной задачи Дирихле (19.21)–(19.22).

Будем искать первую функцию-слагаемое решения u. Предположим, что она может быть представлена в виде произведения двух функций, одна из которых зависит только от переменной x, а вторая — только от переменной y, т.е.

$$u_1(x, y) = X(x) \cdot Y(y).$$

Продифференцировав u_1 дважды по x и по y, имеем:

$$(u_1)'_x = X'Y, \quad (u_1)''_{xx} = X''Y, \quad (u_1)'_y = XY', \quad (u_1)''_{yy} = XY''.$$

Следовательно, при таком представлении u_1 уравнение Лапласа (19.21) приобретает вид

$$X''Y + XY'' = 0.$$

Для ненулевых X и Y (а именно они нас интересуют) последнее уравнение можно записать в форме равенства двух отношений:

$$\frac{X''}{X} = -\frac{Y''}{Y}.$$

Левое отношение здесь может зависеть только от x, правое — только от y. Значит, в силу их равенства, они не зависят ни от x, ни от y, т.е. постоянны.

Положим

$$\frac{X''}{X} = -\frac{Y''}{Y} = -\lambda^2,$$

где λ ($\neq 0$) — некоторый параметр, и вместо одного уравнения, связывающего функции X и Y переменных x и y, будем рассматривать два уравнения:

$$X'' + \lambda^2 X = 0 \quad \text{при} \quad x \in [0, a] \tag{19.25}$$

и

$$Y'' - \lambda^2 Y = 0 \quad \text{при} \quad y \in [-b, b] \tag{19.26}$$

с одним и тем же параметром λ.

Можно считать, что фигурирующее в (19.23) условие $u_1 = 0$ при $x = 0$ и при $x = a$ выполняется за счет функции $X(x)$, т.е. уравнение (19.25) сопровождаем краевыми условиями

$$X(0) = X(a) = 0. \tag{19.27}$$

Составив для (19.25) характеристическое уравнение $r^2 + \lambda^2 = 0$ и найдя его корни $r_{1,2} = \pm \lambda i$, записываем общее решение ОДУ (19.25):

$$X(x) = c_1 \cos \lambda x + c_2 \sin \lambda x. \tag{19.28}$$

Подстановка в него первого из краевых условий (19.27) дает

$$0 = c_1 \cos 0 + c_2 \sin 0 \quad \Rightarrow \quad c_1 = 0.$$

С учетом этого второе условие (19.27) принимает вид

$$0 = c_2 \sin \lambda a. \tag{19.29}$$

Так как ищется нетривиальное решение краевой задачи (19.25), (19.27), то в равенстве (19.29) нельзя полагать $c_2 = 0$. Следовательно, параметр λ может принимать не любые значения, а лишь такие, при которых $\sin \lambda a = 0$, т.е. $\lambda a = n\pi$ $(n \in \mathbf{Z})$. Зафиксировав в (19.28) $\lambda := \frac{n\pi}{a}$, $c_1 := 0$, $c_2 := c$, приходим к тому, что решение краевой задачи (19.25), (19.27) имеет вид

$$X(x) = c \sin \frac{n\pi}{a} x, \tag{19.30}$$

где значения n достаточно считать любыми натуральными, поскольку добавление к ним целых неположительных n новых решений не даст.

Теперь рассматриваем уравнение (19.26) относительно второго сомножителя выражения u_1 — функции $Y = Y(y)$. Здесь корни характеристического уравнения $r^2 - \lambda^2 = 0$ действительны: $r_{1,2} = \pm \lambda$, и, значит, общее решение суть

$$Y(y) = c_1 e^{\lambda y} + c_2 e^{-\lambda y}.$$

В силу соотношений

$$\operatorname{ch} \lambda y = \frac{1}{2}\left(e^{\lambda y} + e^{-\lambda y}\right) \quad \text{и} \quad \operatorname{sh} \lambda y = \frac{1}{2}\left(e^{\lambda y} - e^{-\lambda y}\right)$$

между показательными и гиперболическими функциями, можно перейти к другим произвольным постоянным A и B так, что это

общее решение уравнения (19.26) будет представлено в виде линейной комбинации гиперболических функций:

$$Y(y) = A\operatorname{ch}\lambda y + B\operatorname{sh}\lambda y.$$

Как уже отмечалось, λ — параметр, общий для уравнений (19.25) и (19.26). Поэтому здесь он тоже ограничен значениями вида $\lambda = \frac{n\pi}{a}$. С учетом этого имеем

$$u_1 = X(x)\cdot Y(y) = \left(cA\operatorname{ch}\frac{n\pi}{a}y + cB\operatorname{sh}\frac{n\pi}{a}y\right)\sin\frac{n\pi}{a}x.$$

Согласно построению, такие функции $u_1(x, y)$ удовлетворяют уравнению Лапласа (19.21) при каждом натуральном n, и, ввиду его линейности, ему будут удовлетворять и функции вида

$$\sum_{k=1}^{n}\left(A_k\operatorname{ch}\frac{k\pi}{a}y + B_k\operatorname{sh}\frac{k\pi}{a}y\right)\sin\frac{k\pi}{a}x,$$

а предельным переходом ($n \to \infty$) доказывается, что за u_1 можно принять функцию

$$u_1(x, y) = \sum_{k=1}^{\infty}\left(A_k\operatorname{ch}\frac{k\pi}{a}y + B_k\operatorname{sh}\frac{k\pi}{a}y\right)\sin\frac{k\pi}{a}x. \qquad (19.31)$$

Функция (19.31) удовлетворяет всем выставленным для u_1 требованиям, кроме первых двух краевых условий из (19.23). Чтобы и они были выполнены, нужно соответствующим образом распорядиться двумя семействами параметров A_n и B_n. А именно, замечая, что при фиксированных значениях переменной y выражение (19.31) представляет собой ряд Фурье, функции $\varphi_1(x)$ и $\varphi_2(x)$ также разлагаются в ряд Фурье по функциям $\sin\frac{n\pi}{a}x$ в промежутке $(0, a)$, и из следующих (записанных лишь символически) равенств

$$u_1(x, -b) \quad (\text{ряд Фурье}) \quad = \quad \varphi_1(x) \quad (\text{ряд Фурье}),$$
$$u_1(x, b) \quad (\text{ряд Фурье}) \quad = \quad \varphi_2(x) \quad (\text{ряд Фурье})$$

однозначно находятся коэффициенты A_n, B_n, подстановка которых в (19.31) приводит к единственной искомой функции $u_1(x, y)$.

Функция $u_2(x, y)$ в аддитивном представлении искомого решения $u(x, y)$ данной задачи Дирихле (19.21)–(19.22) может быть получена точно по той же технологии, что и $u_1(x, y)$, рас-

смотрением уравнения Лапласа совместно с граничными условиями (19.24). Но можно поступить и иначе, сделав подходящую замену переменных и частично воспользовавшись уже проделанными при построении u_1 преобразованиями (см. [81]).

Подводя итог этого параграфа, следует сказать, что до появления вычислительной техники было мало альтернатив схематично описанному здесь методу Фурье. Достоинство этого метода — в возможности разделять переменные и сводить граничные и начально-граничные задачи для уравнений с частными производными к аналогичным задачам для обыкновенных дифференциальных уравнений и представлять решения в аналитическом виде. Недостаток — в сложности и, главное, в ограниченности множества задач, для которых удается эффективно выполнить разделение переменных. В основном, такой метод применяется для простых уравнений, заданных на простых областях.

19.4. МЕТОД ПРЯМЫХ

В основу метода прямых могут быть положены две концепции. Одна из них рассматривает этот метод как развитие метода конечных разностей [20, 23, 62, 100, 103]. Другая концепция [23, 138] опирается на идеи проектирования бесконечномерных пространств решений задач для уравнений с частными производными на пространства с конечным функциональным базисом. Остановимся сначала на последнем подходе, применяя его к одной частной задаче, но прежде отметим, что *суть **метода прямых**, обусловившая его название, состоит в том, что, например, в случае уравнения с частными производными относительно функции двух переменных одной из этих переменных придаются постоянные значения* (*иначе, фиксируются некоторые прямые в заданной области изменения переменных*) *и решаются получающиеся при этом соответствующие задачи уже для обыкновенных дифференциальных уравнений.* Таким образом, при решении граничных и начально-граничных задач для УМФ могут быть активно использованы имеющиеся наработки по методам решения начальных и краевых задач для ОДУ. Подобные методы применяют не только при двух, но и при большем числе неизвестных; в таких случаях их называют ***методами плоскостей*** или ***гиперплоскостей*** [100].

Рассмотрим начально-граничную задачу для уравнения теплопроводности (19.18)–(19.20) в немного упрощенной постановке, именно, в случае, когда $\alpha(t) = \beta(t) = 0$, т.е. вместо краевых условий общего вида (19.20) возьмем

$$u(0, t) = 0, \quad u(l, t) = 0 \quad \text{при} \quad t \ge 0. \qquad (19.32)$$

Предположим, что известен набор из n базисных функций

$\varphi_1(x), \ldots, \varphi_n(x)$, которые обладают нужной гладкостью, линейно независимы на отрезке $[0, l]$ и на концах его обращаются в нуль (примеры таких функций приведены в § 17.4). Будем искать приближенное решение данной задачи (19.18), (19.19), (19.32) в виде линейной комбинации

$$u_n(x, t) := \sum_{i=1}^{n} c_i(t)\varphi_i(x) \tag{19.33}$$

выбранных базисных функций с некоторыми пока неопределенными функциональными коэффициентами $c_i(t)$ (опять, как и в предыдущем параграфе, делается своеобразное разделение переменных в решении, только теперь уже в заведомо приближенном). При этом очевидно, что такая функция $u_n(x, t)$ удовлетворяет краевым условиям (19.32).

Имеется несколько способов определения коэффициентов $c_i(t)$ в представлении (19.33) приближенного решения. Один из них основан на ***методе коллокации***, изучавшемся в § 17.4 применительно к краевым задачам для ОДУ. Как и там, на отрезке $[0, l]$ введем сетку $\{x_j\}_{j=1}^{n}$ ***узлов коллокации*** x_j, таких, что

$$0 < x_1 < x_2 < \ldots < x_n < l,$$

и потребуем, чтобы во всех этих узлах функция $u_n(x, t)$ обращала уравнение теплопроводности (19.18) в точное равенство, т.е. чтобы

$$\left.\frac{\partial u_n}{\partial t}\right|_{x=x_j} = a^2 \left.\frac{\partial^2 u_n}{\partial x^2}\right|_{x=x_j} \quad \forall\ j \in \{1, \ldots, n\}. \tag{19.34}$$

Дифференцируя функцию (19.33) по t и дважды по x, получаем

$$\frac{\partial u_n}{\partial t} = \sum_{i=1}^{n} c_i'(t)\varphi_i(x), \quad \frac{\partial^2 u_n}{\partial x^2} = \sum_{i=1}^{n} c_i(t)\varphi_i''(x).$$

Следовательно, условия коллокации (19.34) представляют собой систему n уравнений

$$\begin{cases} \sum_{i=1}^{n} \varphi_i(x_1)c_i'(t) = \sum_{i=1}^{n} a^2\varphi_i''(x_1)c_i(t), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \sum_{i=1}^{n} \varphi_i(x_n)c_i'(t) = \sum_{i=1}^{n} a^2\varphi_i''(x_n)c_i(t). \end{cases} \tag{19.35}$$

Введем n-мерные векторные функции

$$\mathbf{c}(t) := \begin{pmatrix} c_1(t) \\ \vdots \\ c_n(t) \end{pmatrix}, \quad \mathbf{c}'(t) := \begin{pmatrix} c_1'(t) \\ \vdots \\ c_n'(t) \end{pmatrix}$$

и постоянные $n \times n$-матрицы

$$\mathbf{A} := \begin{pmatrix} \varphi_1(x_1) & \cdots & \varphi_n(x_1) \\ \cdots & \cdots & \cdots \\ \varphi_1(x_n) & \cdots & \varphi_n(x_n) \end{pmatrix}, \quad \mathbf{B} := \begin{pmatrix} a^2\varphi_1''(x_1) & \cdots & a^2\varphi_n''(x_1) \\ \cdots & \cdots & \cdots \\ a^2\varphi_1''(x_n) & \cdots & a^2\varphi_n''(x_n) \end{pmatrix}.$$

Тогда система (19.35) коротко записывается в виде

$$\mathbf{A}\mathbf{c}'(t) = \mathbf{B}\mathbf{c}(t). \tag{19.36}$$

В силу линейной независимости базисных функций $\varphi_i(x)$ и несовпадения узлов коллокации x_j (строго внутренних на $[0, l]$), можно рассчитывать на обратимость матрицы $\mathbf{A}$ и, следовательно, на представление уравнения (19.36) нормальной системой ОДУ

$$\mathbf{c}'(t) = \mathbf{A}^{-1}\mathbf{B}\mathbf{c}(t) \tag{19.37}$$

относительно набора $\mathbf{c}(t)$ неизвестных коэффициентов $c_i(t)$ линейной комбинации (19.33).

Для того, чтобы функции $c_i(t)$ из системы (19.37) могли быть определены однозначно, воспользуемся начальным условием (19.19). Поскольку оно должно выполняться в любых точках x отрезка $[0, l]$, это должно иметь место и в узлах коллокации. Следовательно, $u_n(x_j, 0) = \varphi(x_j)$, где $\varphi(x)$ — заданная начальная функция, т.е.

$$\sum_{i=1}^{n} c_i(0)\varphi_i(x_j) = \varphi(x_j), \quad j = 1, \ldots, n. \tag{19.38}$$

Введя n-мерный вектор $\boldsymbol{\varphi} := (\varphi(x_1), \ldots, \varphi(x_n))^T$ и учитывая уже введенные ранее обозначения $\mathbf{c}$ и $\mathbf{A}$, систему начальных условий (19.38) записываем в виде одного векторно-матричного равенства

$$\mathbf{A}\mathbf{c}(0) = \boldsymbol{\varphi}, \quad \text{иначе,} \quad \mathbf{c}(0) = \mathbf{A}^{-1}\boldsymbol{\varphi}. \tag{19.39}$$

Итак, нахождение приближенного решения данной начально-граничной задачи (19.18), (19.19), (19.32) с частными произ-

водными сводится к выбору подходящих семейств базисных функций $\varphi_i(x)$ в представлении приближенного решения (19.33) и к выбору подходящего метода решения задачи Коши (19.37), (19.39) для системы линейных ОДУ первого порядка (с постоянной матрицей), поставляющей функциональные коэффициенты этого представления.

В связи с тем, что в описанном способе решения уравнений в частных производных осуществляют дискретизацию по одной переменной, оставляя другую переменную изменяющейся непрерывно, такие методы называют ***полудискретными***. Однако в большинстве случаев применения метода прямых получающиеся в них системы ОДУ решаются не аналитически, а численно, т. е. фактически производится дискретизация и по другой переменной (что делает название *полудискретный* лишь относительно оправданным).

Ясно, что для повышения точности приближенного решения, получаемого методом прямых, следует увеличивать число n узлов сетки промежутка $[0,l]$ изменения переменной x, т. е. увеличивать число базисных функций $\varphi_i(x)$. Но увеличение n непременно повлечет за собой увеличение числа жесткости g системы ОДУ (19.37) (см. определение 16.5 в § 16.6), что, в свою очередь, делает ее численное решение все более затруднительным. С одной стороны, рост размерности системы (19.37) призывает привлекать для ее решения численные процессы более высоких порядков (чтобы объем вычислений находился в разумных пределах); с другой стороны, одновременное увеличение ее жесткости с ростом n требует применения более устойчивых численных методов (а это, как мы знаем на примере $A(\alpha)$-устойчивых чисто неявных методов, см. § 16.7, равносильно понижению их порядка) и более мелких расчетных шагов по переменной t. Так что, как и во многих других вычислительных ситуациях, здесь важно суметь найти компромисс.

На той же задаче параболического типа (19.18), (19.19), (19.32) рассмотрим теперь конечноразностный подход к построению метода прямых. На отрезке $[0, l]$ зафиксируем равноотстоящие точки $x_i = ih$ $(i = 0, 1, \ldots, n)$ с шагом $h = \dfrac{l}{n}$ и будем искать приближенные решения $u(x_i, t)$ данной задачи на определяемых этими точками прямых $x = x_i$ $(i = 1, 2, \ldots, n-1)$ в области

$t > 0$ (рис. 19.2).

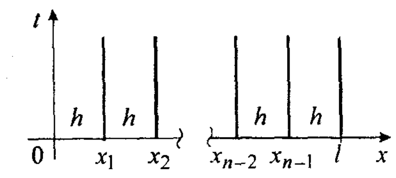


Рис. 19.2. *Прямые, на которых ищутся приближенные решения задачи* (19.18), (19.19), (19.32) *полудискретным методом прямых*

Замораживая в уравнении теплопроводности (19.18) переменную t, ко второй производной по переменной x применим простейшую формулу симметричной аппроксимации (13.20):

$$\frac{\partial^2 u(x,t)}{\partial x^2} = \frac{u(x-h,t) - 2u(x,t) + u(x+h,t)}{h^2} + O(h^2). \qquad (19.40)$$

Подстановка формулы (19.40) в правую часть уравнения (19.18) при $x = x_i$ $(i = 1, 2, \ldots, n-1)$ дает $n-1$ уравнений вида

$$\frac{du(x_i,t)}{dt} = \frac{a^2}{h^2}\left[u(x_{i-1},t) - 2u(x_i,t) + u(x_{i+1},t)\right] + O(h^2).$$

Отбросим в них слагаемые $O(h^2)$ и учтем в первом и в последнем уравнениях краевые условия (19.32), согласно которым

$$u(x_0,t) = 0, \quad u(x_n,t) = 0.$$

Тогда, вводя для простоты функции переменной t

$$v_i(t) :\approx u(x_i,t), \qquad (19.41)$$

приходим к системе обыкновенных дифференциальных уравнений

$$\begin{cases} \dfrac{dv_1(t)}{dt} = \dfrac{a^2}{h^2}[-2v_1(t) + v_2(t)], \\ \dfrac{dv_i(t)}{dt} = \dfrac{a^2}{h^2}[v_{i-1}(t) - 2v_i(t) + v_{i+1}(t)], \quad i = 2, \ldots, n-2, \\ \dfrac{dv_{n-1}(t)}{dt} = \dfrac{a^2}{h^2}[v_{n-2}(t) - 2v_{n-1}(t)] \end{cases} \qquad (19.42)$$

с начальными условиями

$$v_i(0) \quad (= u(x_i,0)) = \varphi(x_i), \quad i = 1, 2, \ldots, n-1, \qquad (19.43)$$

получающимися из заданного начального условия (19.19). Коротко эту систему ОДУ можно записать в векторно-матричном виде

$$\mathbf{v}' = \mathbf{A}\mathbf{v}, \qquad (19.42а)$$

где

$$\mathbf{A} := \frac{a^2}{h^2}\begin{pmatrix} -2 & 1 & 0 & \cdots & 0 & 0 \\ 1 & -2 & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 & -2 \end{pmatrix}, \quad \mathbf{v} := \begin{pmatrix} v_1(t) \\ v_2(t) \\ \cdots \\ v_{n-1}(t) \end{pmatrix}, \quad \mathbf{v}' = \begin{pmatrix} v_1'(t) \\ v_2'(t) \\ \cdots \\ v_{n-1}'(t) \end{pmatrix},$$

причем $(n-1)\times(n-1)$-матрица $\mathbf{A}$ симметрична и имеет трехдиагональную структуру.

Не составит большого труда получить аналитический вид решения системы (19.42) рассматриваемого конечноразностного метода прямых. С этой целью введем числа

$$\lambda_i := -\frac{a^2}{h^2}\left(2 - 2\cos\frac{i}{n}\pi\right), \quad i = 1, 2, \ldots, n-1 \qquad (19.44)$$

и векторы

$$\mathbf{s}_i := \left(\sin\frac{i}{n}\pi; \sin\frac{2i}{n}\pi; \ldots; \sin\frac{n-1}{n}i\pi\right)^T, \quad i = 1, 2, \ldots, n-1 \qquad (19.45)$$

и убедимся, что при одинаковых значениях i они образуют собственные пары матрицы $\mathbf{A}$ системы (19.42а).

Действительно, так как

$$\lambda_i \mathbf{s}_i = -\frac{2a^2}{h^2}\begin{pmatrix} \sin\frac{i}{n}\pi - \sin\frac{i}{n}\pi\cdot\cos\frac{i}{n}\pi \\ \sin\frac{2i}{n}\pi - \sin\frac{2i}{n}\pi\cdot\cos\frac{i}{n}\pi \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \sin\frac{n-1}{n}i\pi - \sin\frac{n-1}{n}i\pi\cdot\cos\frac{i}{n}\pi \end{pmatrix}$$

и

$$\mathbf{A}\mathbf{s}_i = \frac{a^2}{h^2}\begin{pmatrix} -2\sin\frac{i}{n}\pi + \sin\frac{2i}{n}\pi \\ \sin\frac{i}{n}\pi - 2\sin\frac{2i}{n}\pi + \sin\frac{3i}{n}\pi \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \sin\frac{n-2}{n}i\pi - 2\sin\frac{n-1}{n}i\pi \end{pmatrix} =$$

$$= \frac{a^2}{h^2}\begin{pmatrix} -2\sin\frac{i}{n}\pi + 2\sin\frac{i}{n}\pi\cdot\cos\frac{i}{n}\pi \\ -2\sin\frac{2i}{n}\pi + 2\sin\frac{2i}{n}\pi\cdot\cos\frac{i}{n}\pi \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ -2\sin\frac{n-1}{n}i\pi + \sin\frac{n-2}{n}i\pi + \sin\frac{n}{n}i\pi \end{pmatrix} =$$

$$= -\frac{2a^2}{h^2}\begin{pmatrix} \sin\frac{i}{n}\pi - \sin\frac{i}{n}\pi\cdot\cos\frac{i}{n}\pi \\ \sin\frac{2i}{n}\pi - \sin\frac{2i}{n}\pi\cdot\cos\frac{i}{n}\pi \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \sin\frac{n-1}{n}i\pi - \sin\frac{n-1}{n}i\pi\cdot\cos\frac{i}{n}\pi \end{pmatrix},$$

то

$$\mathbf{A}\mathbf{s}_i = \lambda_i \mathbf{s}_i \quad \forall i \in \{1, 2, \ldots, n-1\}.$$

Анализируя собственные числа (19.44) матрицы $\mathbf{A}$, видим, что они отрицательны и различны. Следовательно, общее решение однородной системы ОДУ (19.42а) через эти числа и отвечающие им собственные векторы (19.45) можно представить в виде

$$\mathbf{v}(t) = c_1 e^{\lambda_1 t}\mathbf{s}_1 + c_2 e^{\lambda_2 t}\mathbf{s}_2 + \ldots + c_{n-1}e^{\lambda_{n-1}t}\mathbf{s}_{n-1}. \qquad (19.46)$$

Для выделения из него нужного частного решения, т.е. для нахождения конкретных значений произвольных постоянных $c_1, c_2, \ldots, c_{n-1}$, привлекаем начальные условия (19.43), составляя с их помощью линейную алгебраическую систему относительно искомых значений.

Можно воспользоваться и готовой формулой представления решения $\tilde{\mathbf{v}}(t)$ задачи Коши (19.42)–(19.43) [23]:

$$\tilde{\mathbf{v}}(t) = \mathbf{S}^{-1}diag(e^{\lambda_i t})\mathbf{S}\mathbf{v}(0),$$

где $\mathbf{S} := (\mathbf{s}_1; \mathbf{s}_2; \ldots; \mathbf{s}_{n-1})$ — квадратная матрица, столбцами которой служат собственные векторы (19.45), а

$$\mathbf{v}(0) := (v_1(0); v_2(0); \ldots; v_n(0))^T$$

— вектор значений начальной функции $\varphi(x)$ в узлах x_i (соответствующий условиям (19.43)).

Пример 19.1. Дана задача

$$\frac{\partial u}{\partial t}=\frac{\partial^2 u}{\partial x^2}, \quad u(x,0)=\sqrt{x(1-x)}, \quad x\in[0,1], \qquad (19.47)$$
$$u(0,t)=0, \quad u(1,t)=0, \quad t\ge 0.$$

Рассмотрим процесс построения ее приближенных решений методом прямых в двух показанных выше версиях: конечномерной и проекционной. При этом ограничимся самым простейшим случаем, когда берется всего одна прямая, осуществляющая сечение области $[0,1]\times[0,\infty)$ задания переменных x и t, а именно, прямая $x=x_1:=0.5$.

I способ (*конечноразностный*). Воспользуемся представлением (19.46) приближенного решения $v_1(t)$ на прямой $x=0.5$. Учитывая, что задача (19.47) есть частный случай задачи (19.18), (19.19), (19.32) с $a=1$, $l=1$, $\varphi(x)=\sqrt{x(1-x)}$, и полагая $n=2$, $h=0.5$, по формулам (19.44), (19.45) находим

$$\lambda_1=-\frac{1}{0.25}\left(2-2\cos\frac{\pi}{2}\right)=-8, \qquad s_1=\sin\frac{\pi}{2}=1;$$

тогда, согласно (19.46),

$$v_1(t)=c_1e^{\lambda_1 t}s_1=c_1e^{-8t} \qquad (19.48)$$

— общее решение системы типа (19.42), выродившейся в одно уравнение

$$\frac{dv_1}{dt}=-8v_1.$$

Из начального для него условия типа (19.43)

$$v_1(0)=u(x_1,0)=\sqrt{0.5(1-0.5)}=0.5$$

подстановкой в (19.48) $t=0$ получаем $c_1=0.5$, и, следовательно, искомым приближенным решением задачи (19.47) на прямой $x=0.5$ можно считать функцию

$$\widetilde{v}_1(t)=0.5e^{-8t}.$$

II способ (*проекционный*). В соответствии с формулой (19.33) при $n=1$ ищем приближенное решение в виде функции $u_1(x,t)=c_1(t)\varphi_1(x)$. В качестве базисной функции $\varphi_1(x)$ возьмем первую функцию из семейства функций $\varphi_i(x)=x^i(1-x)$, удовлетворяющих данным однородным краевым условиям. Дважды дифференцируя выбранную функцию $\varphi_1(x)=x(1-x)$, получаем $\varphi_1''(x)\equiv -2$. Таким образом, в роли $\mathbf{A}$ в уравнении (19.36) выступает $\varphi_1(x_1)=0.5(1-0.5)=0.25$, а в роли $\mathbf{B}$ — $a^2\varphi_1''(x_1)=-2$. Значит, в представлении приближенного решения

$$u_1(x,t)=x(1-x)c_1(t) \qquad (19.49)$$

функция $c_1(t)$ должна находиться как решение уравнения

$$0.25c_1'(t)=-2c_1(t) \quad \text{или, проще,} \quad c_1'(t)=-8c_1(t).$$

Присовокупив к этому дифференциальному уравнению начальное условие типа (19.39), т.е.

$$c_1(0)=4\sqrt{0.5(1-0.5)}=2,$$

находим решение полученной задачи Коши для ОДУ:

$$c_1(t)=2e^{-8t}.$$

Подстановкой этого выражения $c_1(t)$ в (19.49) получаем приближенное решение

$$u_1(x,t)=2x(1-x)e^{-8t}$$

(при $x=0.5$ совпадающее с решением $\widetilde{v}_1(t)$, найденным предыдущим способом).

Если высказанное несколькими страницами ранее суждение о возрастании жесткости систем ОДУ (19.37) с ростом их размерности в методе прямых проекционной природы, не было достаточно обосновано, то для конечноразностного метода прямых, сводящегося к решению систем ОДУ вида (19.42), число жесткости может быть точно подсчитано.

Действительно, в силу вещественности собственных чисел (19.44) матрицы $\mathbf{A}$ системы (19.42а), определенное в § 16.6 число жесткости g системы ОДУ, рассматриваемое как функция размерности $n-1$ этой системы, есть

$$g(n-1)=\frac{\max\limits_{i\in\{1,\ldots,n-1\}}\{|\lambda_i|\}}{\min\limits_{i\in\{1,\ldots,n-1\}}\{|\lambda_i|\}}=\frac{\frac{a^2}{h^2}\left(2-2\cos\frac{n-1}{n}\pi\right)}{\frac{a^2}{h^2}\left(2-2\cos\frac{1}{n}\pi\right)}=$$
$$=\frac{1+\cos\frac{\pi}{n}}{1-\cos\frac{\pi}{n}}=\frac{\cos^2\frac{\pi}{2n}}{\sin^2\frac{\pi}{2n}}.$$

Так как при $n\to\infty$ имеет место эквивалентность бесконечно малых $\sin\frac{\pi}{2n}$ и $\frac{\pi}{2n}$, значит, $g(n-1)\xrightarrow[n\to\infty]{}\infty$ со скоростью $O(n^2)$, что говорит о быстром ухудшении обусловленности систем с ростом n.

В данном конкретном случае применения конечноразностного метода прямых жесткость получающихся систем ОДУ не играет роли, поскольку для этих систем известны аналитические решения. Однако следует иметь в виду, что здесь намеренно рассматривалась простая задача, чтобы не затенять суть метода и

выявить его характерные черты. На самом деле область приложения метода прямых весьма обширна, причем употребляется он чаще всего как сугубо численный метод, где явление жесткости уже нельзя не учитывать. Особенности при применении метода прямых к уравнениям в частных производных различных типов имеются, но они не столь принципиальны. Более подробно об этом методе, его точности и разновидностях см., например, в [20, 62, 100 и др.].

19.5. ВАРИАЦИОННЫЕ МЕТОДЫ. МЕТОД РИТЦА (ОБЩАЯ СХЕМА)

В гл. 17, посвященной краевым задачам для ОДУ, мы лишь упомянули о существовании метода Ритца, реализующего вариационный подход к построению приближенно-аналитических решений (см. замечание 17.6 в § 17.5), отдав предпочтение освещению метода Галёркина, во многих случаях приводящего в итоге к тем же результатам. Здесь попытаемся вкратце отразить основные моменты, касающиеся идеологии вариационных методов и, в частности, метода Ритца, служащего, как и метод Галёркина, фундаментом для метода конечных элементов (применительно к ОДУ последний уже рассматривался в § 17.6). При этом, как и при описании метода Галёркина, сначала поднимемся на уровень функционального анализа.

Пусть требуется найти приближенное решение уравнения

$$Ly = f, \tag{19.49}$$

где $L\colon\ D(L) \subseteq H \to H$ — линейный оператор, H — вещественное гильбертово пространство (в общем случае бесконечномерное). Предполагаем, что это операторное уравнение имеет в $D(L)$ единственное решение y^*, что обеспечивается, например, требованием положительности оператора L в смысле выполнения условия [20, 126 и др.]

$$(Ly, y) \ge 0 \quad \forall y \in D(L), \quad \text{причем} \quad (Ly, y) = 0 \Leftrightarrow y = 0.$$

Данному уравнению (19.49) сопоставляется какой-нибудь функционал $J[y]: H \to \mathbf{R}_1$ такой, что он ограничен снизу и достигает своего минимального значения J_* в единственной точке, совпадающей с искомым решением y^* уравнения (19.49). Тогда исходная задача в постановке (19.49) сводится к решению экстремальной задачи

$$J[y] \to \min, \quad y \in D(L). \tag{19.50}$$

Можно сказать, что здесь осуществляется подход, который характерен для классического вариационного исчисления [190, 196], где как раз наоборот, поиск экстремалей функционалов сводится к решению уравнений в вариациях функционалов (аналогах дифференциалов функций).

Разнообразие методов, опирающихся на сведение задач вида (19.49) к задачам вида (19.50) и называемых ***вариационными методами***, очевидно, связано и с тем, что можно конструировать разные функционалы $J[y]$, обладающие нужными свойствами, и с тем, что можно по-разному понимать близость приближенных решений $\tilde{y}$ задачи (19.50) к точному решению y^* (например, или в смысле малости нормы $\|y^* - \tilde{y}\|$, или в смысле малости разности $J[\tilde{y}] - J[y^*]$, или еще в каком-либо смысле), и с тем, что можно по-разному строить процессы получения приближений $\tilde{y}$ к y^*.

Если оператору L в (19.49) сопоставляется функционал $J[y]$ в (19.50) вида

$$J[y] := \|Ly - f\|^2 = (Ly, Ly) - 2(Ly, f) + \|f\|^2, \tag{19.51}$$

т.е. функционал, который естественно назвать квадратичным, то соответствующий вариационный метод называется ***методом наименьших квадратов****). Совершенно очевидно, что такой функционал всегда имеет минимум на множестве $D(L)$, и в случае, когда уравнение (19.49) имеет единственное решение $y^* \in D(L)$, то это решение

$$y^* = \arg\min_{y \in D(L)} J[y], \quad \text{причем} \quad J[y^*] = J_* = 0.$$

Другим часто используемым в (19.50) функционалом является так называемый ***функционал энергии*** [47]:

$$J[y] := (Ly, y) - 2(y, f). \tag{19.52}$$

Взаимно однозначное соответствие между решениями задачи (19.50) с $J[y]$ вида (19.52) и задачи (19.49) устанавливается при более жестких требованиях к оператору L, чем в случае использования функционала (19.51), в частности, при условии симметричности и положительной определенности оператора L

*) К минимизации таких функционалов в конечномерных пространствах, представлявших там собою просто функции нескольких переменных, мы уже приходили ранее при решении разных задач: при решении систем конечномерных нелинейных уравнений (§ 7.6), при решении переопределенных СЛАУ (§ 10.1), при подборе параметров эмпирических функций (§ 10.1), при построении обобщенных многочленов наилучших среднеквадратических приближений (§ 10.2).

(т.е. если

$$(Lu, v) = (u, Lv) \quad \forall\, u, v \in D(L)$$

и, кроме того,

$$\exists\, \gamma > 0: \quad (Lu, u) \ge \gamma \|u\|^2 \quad \forall\, u \in D(L) \;).$$

Вариационный метод, основанный на минимизации функционала (19.52), называют ***энергетическим методом*** [47, 126].

Заметим, что часто вместо функционала энергии вида (19.52) (иначе, ***функционала энергетического метода*** [126]) в вариационной задаче (19.50) можно встретить функционал более общего вида [20, 125, 148, 171]

$$J[y] := (Ly, y) - (y, f) - (f, y),$$

определенный в комплексных гильбертовых пространствах (т.е. в пространствах, для которых $(y, f) = \overline{(f, y)}$) и совпадающий с (19.52) в случае вещественных гильбертовых пространств.

Также следует отметить наличие связи между энергетическим методом и методом наименьших квадратов: при определенных условиях уравнение (19.49) равносильно уравнению

$$L^* Ly = L^* f$$

(где L^* — оператор, сопряженный к оператору L), решение которого энергетическим методом означает минимизацию функционала

$$\left(L^* Ly, y\right) - 2\left(L^* f, y\right) = \|Ly - f\|^2 - \|f\|^2,$$

который лишь постоянным слагаемым $-\|f\|^2$ отличается от функционала (19.51) метода наименьших квадратов.

Не вдаваясь в детали, существенные при решении вопросов об одновременном существовании и единственности решений исходной задачи (19.49) и ее вариационной формулировки (19.50), примем за основу энергетический метод и схематично рассмотрим один из способов построения приближенных решений задачи минимизации (19.50), связываемый с именем В.Ритца*), предложившего этот способ в 1908 г.

В условиях симметричности и положительности оператора L, требуемых, как указано выше, для применения энергетического метода, на элементах множества $D(L)$ определяется ***энергетическое скалярное произведение***

$$[u, v] := (Lu, v) \quad \forall\, u, v \in D(L)$$

---

*) Ритц Вальтер (1878–1909) —физик и математик. Родился и учился в Швейцарии, продолжил обучение и далее работал в университетах Германии.

и, соответственно, ***энергетическая норма***

$$\| u \| := \sqrt{[u, u]} = \sqrt{(Lu, u)}.$$

В этой метрике множество $D(L)$, ставшее унитарным пространством, дополняется до полного, и построенное таким образом гильбертово пространство H_L называется ***энергетическим пространством***; при этом доказывается, что если $u \in H_L$, то и $u \in H$.

Выберем в пространстве H_L последовательность элементов

$$\varphi_1, \varphi_2, \ldots \varphi_n, \ldots, \tag{19.53}$$

называемых ***координатными элементами***, таких, что:

1) при любом n элементы $\varphi_1, \varphi_2, \ldots \varphi_n$ линейно независимы;

2) система элементов (19.53) ***полна по энергии***, т.е.

$$\forall u \in H_L \ \forall \varepsilon > 0 \ \exists n_0 \in \mathbf{N} \ \exists \{\alpha_1, \ldots, \alpha_{n_0} | \alpha_k \in \mathbf{R}_1\}: \ \left\| u - \sum_{k=1}^{n_0} \alpha_k \varphi_k \right\| < \varepsilon.$$

По ***методу Ритца*** приближенное решение вариационной задачи (19.50) с функционалом (19.52), а значит (в оговоренных условиях), и исходной задачи (19.49), ищется в виде

$$y_n = \sum_{i=1}^{n} c_i \varphi_i\,, \tag{19.54}$$

где числа $c_i \quad (i = 1, 2, \ldots, n)$ — коэффициенты линейной комбинации n первых координатных элементов последовательности (19.53) в представлении (19.54) — подбираются такими, при которых

$$J[y_n] = \min_{c_i \in \mathbf{R}_n} J[\sum_{i=1}^{n} c_i \varphi_i].$$

Следовательно, задача минимизации функционала энергии (19.52) сводится к задаче минимизации функции n переменных

$$F(c_1, \ldots, c_n) := J[\sum_{i=1}^{n} c_i \varphi_i] = \left(L\sum_{i=1}^{n} c_i \varphi_i, \ \sum_{i=1}^{n} c_i \varphi_i \right) - 2\left(\sum_{i=1}^{n} c_i \varphi_i, \ f \right), \tag{19.55}$$

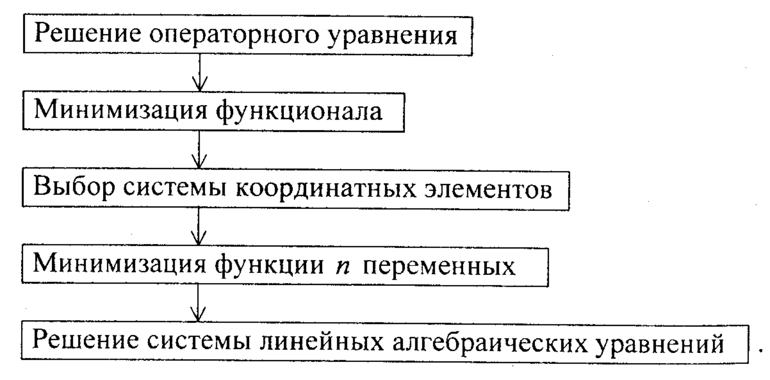
которая, в свою очередь, с учетом предъявленных к линейному оператору L требований заменяется равносильной задачей решения СЛАУ

$$\begin{cases} (L\varphi_1, \varphi_1)c_1 + (L\varphi_1, \varphi_2)c_2 + \ldots + (L\varphi_1, \varphi_n)c_n = (\varphi_1, f), \\ (L\varphi_2, \varphi_1)c_1 + (L\varphi_2, \varphi_2)c_2 + \ldots + (L\varphi_2, \varphi_n)c_n = (\varphi_2, f), \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ (L\varphi_n, \varphi_1)c_1 + (L\varphi_n, \varphi_2)c_2 + \ldots + (L\varphi_n, \varphi_n)c_n = (\varphi_n, f). \end{cases} \tag{19.56}$$

Эта система*), называемая ***системой Ритца***, получается приравниванием к нулю производных $\dfrac{\partial F(c_1, \dots, c_n)}{\partial c_k}$ $(k = 1, 2, \dots, n)$, что обеспечивает выполнение необходимых, а в данном случае и достаточных условий минимума функции (19.55); ее однозначная разрешимость вытекает из линейной независимости координатных элементов φ_i.

Улучшение качества приближенных решений y_n (19.54), получаемых процессом Ритца, (а если говорить более строго, невозрастание энергетических норм $\|y^* - y_n\|$ [126]) с ростом числа n используемых координатных элементов из (19.53) есть следствие сепарабельности исходного гильбертова пространства H и второго требования к системе координатных элементов, т.е. ее полноты.

Итак, вариационный метод Ритца можно уложить в следующую схему связи задач, порождающих одна другую:



.

При этом некоторые промежуточные задачи нужны, фактически, лишь для понимания того, откуда берется следующая. Собственно, для формального получения приближений по методу Ритца требуется лишь выбрать подходящие элементы φ_i в представлении (19.54) и решить систему (19.56) с симметричной положительно определенной матрицей, чтобы найти оптимальные коэффициенты этого представления y_n (разумеется в предположении, что уже есть вариационная постановка исходной задачи).

---

*) Такой вид она имеет в предположении, что все $\varphi_i \in D(L)$. В более общем случае $\varphi_i \in H_L$ в системе (19.56) вместо $(L\varphi_i, \varphi_j)$ фигурируют энергетические скалярные произведения $[\varphi_i, \varphi_j]$, что связано с тонкостями процедуры пополнения пространства $D(L)$.

Полное совпадение системы Ритца (19.56) с системой (17.59)–(17.60) метода Галеркина для нахождения коэффициентов c_i приближенного решения вида (17.57), идентичного (19.54), для операторного уравнения (17.56) того же, что и (19.49), говорит о близости методов вариационной и проекционной природы. При этом можно указать литературные источники, где метод Ритца вписывается в теорию проекционных методов [47, 148], но иногда наоборот, проекционные методы такие, как метод Галеркина или более общий метод моментов, называют вариационными [117]; параллельное изучение этих методов см. в [81]. Считается, что метод Ритца имеет более узкую сферу применения, поскольку предъявляет более жесткие требования к оператору L задачи (19.49), чем метод Галеркина, но имеет перед последним то преимущество, что в рамках этого метода возможно получение более эффективных оценок погрешностей приближенных решений.

19.6. МЕТОД РИТЦА ДЛЯ ДВУМЕРНОЙ ЗАДАЧИ ДИРИХЛЕ

Естественной сферой приложения энергетического метода и, в частности, метода Ритца, общее представление о которых дано в предыдущем параграфе, являются краевые задачи для самосопряженных уравнений эллиптического типа. Чтобы понять, как реализуется метод Ритца на конкретных задачах, рассмотрим одну из наиболее простых подходящих для этого случая задач, а именно, задачу Дирихле для уравнения Пуассона (19.5) в ограниченной открытой области Ω точек $(x; y) \in \mathbf{R}_2$ с кусочно-гладкой границей $\mathbf{\Gamma}$ и однородными краевыми условиями. При этом для упрощения привязки этой задачи к вариационной формулировке, умножим уравнение (19.5) на -1, и вместо $-f(x, y)$ будем снова использовать $f(x, y)$.

Итак, пусть требуется вариационным методом найти приближенное решение задачи

$$\begin{cases} -\dfrac{\partial^2 u}{\partial x^2} - \dfrac{\partial^2 u}{\partial y^2} = f(x, y) & \text{при} \quad (x; y) \in \Omega, \qquad (19.57) \\ u(x, y) = 0 & \text{при} \quad (x; y) \in \mathbf{\Gamma}, \qquad (19.58) \end{cases}$$

где функция $f(x, y)$ предполагается суммируемой с квадратом в области Ω.

Сопоставляя задачу с общей постановкой в операторной форме (19.49), примем за исходное пространство $L_2(\Omega)$ функций двух переменных $w(x, y)$, суммируемых в Ω с квадратом (как известно, гильбертово [45, 148, 171 и др.]), а оператором L, действующим в этом пространстве, будем считать оператор Лап-

ласа, взятый с противоположным знаком, т.е. полагаем, что в уравнении (19.49), примененном к данному случаю*),

$$Lu := -\Delta u = -\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right). \qquad (19.59)$$

Тогда областью определения $D(L)$ введенного равенством (19.59) оператора L является сужение пространства H на множество таких функций, которые обладают в области Ω нужной гладкостью (например, дважды непрерывно дифференцируемы) и на ее границе Γ обращаются в нуль.

Так как скалярное произведение в пространстве $L_2(\Omega)$ есть

$$(w, v) = \iint_{\Omega} wv d\Omega = \iint_{\Omega} w(x, y) v(x, y) dxdy, \qquad (19.60)$$

то в соответствии с (19.59)

$$(Lu, v) = -\iint_{\Omega} v\Delta u d\Omega = -\iint_{\Omega} v\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) dxdy. \qquad (19.61)$$

Пользуясь этим, покажем симметричность и положительность оператора L, что, согласно § 19.5, нужно для эквивалентной вариационной постановки задачи (19.57)–(19.58).

Для установления факта симметрии L рассмотрим следующую разность скалярных произведений вида (19.61) на множестве функций из $D(L)$:

$$(Lu, v) - (Lv, u) = -\iint_{\Omega} v\Delta u d\Omega + \iint_{\Omega} u\Delta v d\Omega =$$

$$= -\iint_{\Omega}\left[v\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) - u\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right)\right] dxdy = -\iint_{\Omega} T(x, y) dxdy,$$

где

$$T(x, y) := v\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) - u\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right) =$$

$$= \frac{\partial}{\partial x}\left(v\frac{\partial u}{\partial x} - u\frac{\partial v}{\partial x}\right) + \frac{\partial}{\partial y}\left(v\frac{\partial u}{\partial y} - u\frac{\partial v}{\partial y}\right)$$

(последнее равенство легко проверяется справа налево). Далее

*) Роль искомого элемента y в обозначениях предыдущего параграфа выполняет $u = u(x, y)$, поскольку для y здесь традиционно отведено место второй независимой переменной.

применяем известную ***формулу Грина***, согласно которой

$$\iint_{\Omega}\left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dxdy = \int_{\Gamma} Pdx + Qdy, \qquad (19.62)$$

полагая в данном случае

$$Q := v\frac{\partial u}{\partial x} - u\frac{\partial v}{\partial x}, \quad P := -v\frac{\partial u}{\partial y} + u\frac{\partial v}{\partial y},$$

и получаем

$$(Lu, v) - (Lv, u) = -\int_{\Gamma}\left(v\frac{\partial u}{\partial y} + u\frac{\partial v}{\partial y}\right) dx - \left(v\frac{\partial u}{\partial x} - u\frac{\partial v}{\partial x}\right) dy.$$

Так как функции u и v обращаются в нуль на $\mathbf{\Gamma}$, то последний интеграл равен нулю, т.е. $(Lu, v) = (Lv, u)$, что говорит о симметрии L.

Теперь покажем положительность оператора L. По формуле (19.61) имеем

$$(Lu, u) = -\iint_{\Omega} u\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) dxdy = -\iint_{\Omega} S(x, y) dxdy,$$

где, очевидно, справедливо следующее выражение подынтегральной функции:

$$S(x, y) := u\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = \frac{\partial}{\partial x}\left(u\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(u\frac{\partial v}{\partial y}\right) - \left(\frac{\partial u}{\partial x}\right)^2 - \left(\frac{\partial u}{\partial y}\right)^2.$$

Это позволяет преобразовать скалярное произведение (Lu, u) (опять привлекая формулу Грина (19.62)) так:

$$(Lu, u) = -\iint_{\Omega}\left[\frac{\partial}{\partial x}\left(u\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(u\frac{\partial u}{\partial y}\right)\right] dxdy + \iint_{\Omega}\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] dxdy =$$

$$= -\int_{\Gamma}\left[\left(-u\frac{\partial u}{\partial x}\right) dx + \left(u\frac{\partial u}{\partial y}\right) dy\right] + \iint_{\Omega}\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] dxdy. \quad (19.63)$$

Криволинейный интеграл в последнем выражении равен нулю, в силу (19.58), а двойной интеграл неотрицателен, в силу неотрицательности подынтегральной функции; следовательно, $(Lu, u) \geq 0$. Так как

$$\iint_{\Omega}\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right] dxdy = 0 \quad \Leftrightarrow \quad \begin{cases} \frac{\partial u}{\partial x} = 0, \\ \frac{\partial u}{\partial y} = 0, \end{cases}$$

функция $u(x, y) \equiv const$ при любых $(x; y) \in \Omega$, а в силу ее непрерывности, при любых $(x; y) \in \overline{\Omega} := \Omega \cup \mathbf{\Gamma}$. Но на границе $\mathbf{\Gamma}$ функция u равна нулю по условию; значит, она равна нулю во всей области $\overline{\Omega}$, т.е.

$$(Lu, u) = 0 \quad \Leftrightarrow \quad u = 0.$$

Итак, введенный равенством (19.59) оператор L симметричен и положителен в Ω, что позволяет свести решение задачи (19.57)–(19.58) к минимизации функционала вида (19.52). Поскольку предыдущие преобразования, приводящие к формуле (19.63), показали, что для определяемых постановкой задачи функций u справедливо выражение

$$(Lu, u) = \iint\limits_{\Omega}\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2\right]dxdy \tag{19.64}$$

(вошедшее в литературу как ***интеграл Дирихле*** [62, 81, 165]), функционал энергии, подлежащий минимизации, есть

$$\Phi[u] := \iint\limits_{\Omega}\left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2 - 2uf(x, y)\right]dxdy. \tag{19.65}$$

В общем, для последующего построения приближений к решению данной задачи методом Ритца он и не нужен, так как для этого достаточно воспользоваться заготовками предыдущего параграфа, а вот вид этого функционала, не содержащего вторых производных, говорит о возможности получения этим методом решений (в обобщенном смысле) меньшей гладкости, чем гармонические функции.

Ставя цель получить приближенные решения данной задачи Дирихле методом Ритца, т.е. найти при некоторых фиксированных $n \in \mathbf{N}$ функции вида

$$u_n(x, y) = \sum_{i=1}^{n} c_i \varphi_i(x, y) \tag{19.66}$$

(в соответствии с представлением (19.54)), которые в идеале при $n \to \infty$ должны доставлять минимум функционалу (19.65), следует понимать, что во многом успех зависит от того, насколько удачно выбрана последовательность координатных функций $\varphi_i = \varphi_i(x, y)$, $i \in \mathbf{N}$. При их подборе, как правило, приходится ориентироваться на конкретный вид области Ω (которая для обоснованного применения такого метода должна быть достаточно простой), по возможности, учитывать какие-то свойства решения (например, четность), проверять систему $\{\varphi_i\}$ на линейную независимость и полноту. На этот счет имеются некоторые общие соображения. Например, в [81] доказывается, что если в области $\overline{\Omega}$ задать непрерывную функцию $\omega = \omega(x, y)$, такую, чтобы в Ω она имела непрерывные производные первого порядка и, кроме того, была положительна в области Ω и обращалась в нуль на ее границе $\mathbf{\Gamma}$, то последовательность функций вида

$$\varphi_0 = \omega, \quad \varphi_1 = \omega x, \quad \varphi_2 = \omega y, \quad \varphi_3 = \omega x^2, \quad \varphi_4 = \omega xy, \ \ldots, \tag{19.67}$$

представляющих собой произведения заданной функции $\omega(x, y)$ на произведения всевозможных степеней переменных x и y, отвечает требуемым свойствам линейной независимости и полноты в множестве $D(L)$, где задан оператор Лапласа, с метрикой пространства $L_2(\Omega)$. В качестве функции $\omega(x,y)$, например, для прямоугольной области $[-a, a] \times [-b, b]$ может быть взята функция

$$\omega(x, y) := (x^2 - a^2)(y^2 - b^2); \tag{19.68}$$

для круга радиуса r с центром в начале координат —

$$\omega(x, y) := r^2 - x^2 - y^2.$$

Наряду с полиномиальными координатными функциями в методе Ритца применяют также тригонометрические функции [125]. Так, в случае прямоугольника $[0, a] \times [0, b]$ полной системой линейно независимых функций, пригодных для поиска приближенных решений в виде (19.66), является последовательность функций, порождаемая формулой [47]

$$\varphi_{kj}(x, y) = \sin\left(k\frac{\pi x}{a}\right)\sin\left(j\frac{\pi y}{b}\right), \quad k, j \in \mathbf{N}.$$

В случае сложного контура границы $\mathbf{\Gamma}$ области Ω при построении системы координатных функций применяют технику ***R-функций*** (***функций В.Л.Рвачева***), основанную на использовании алгебры логики [143].

Если подходящие координатные функции φ_i для выражения (19.66) выбраны, остается: подсчитать коэффициенты линейной алгебраической системы (типа 19.56)

$$\sum_{j=1}^{n} a_{ij} c_j = b_i, \quad i = 1, 2, \ldots, n \tag{19.69}$$

по формулам (соответствующим выражениям скалярных произведений (19.61), (19.60))

$$a_{ij} := (L\varphi_i, \varphi_j) = -\iint\limits_{\Omega}\varphi_j\left(\frac{\partial^2 \varphi_i}{\partial x^2} + \frac{\partial^2 \varphi_i}{\partial y^2}\right)dxdy \tag{19.70}$$

(при подсчете a_{ij} можно учесть (19.64)),

$$b_i := (\varphi_i, f) = \iint_\Omega \varphi_i f(x, y)dxdy, \qquad (19.71)$$

найти из системы (19.69) числа c_i $(i = 1, 2, \ldots, n)$ и подставить их вместе с выбранными функциями φ_i в форму приближенного решения (19.66).

Пример 19.2. Методом Ритца найдем приближение $u_2(x, y)$ к решению $u(x, y)$ задачи Дирихле для уравнения Пуассона

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -x - y \qquad (19.72)$$

в квадрате $[-1, 1]\times[-1, 1]$ при условии, что $u(x, y) = 0$ на его границе.

В соответствии с одной из приведенных выше рекомендаций по выбору координатных функций зададимся функцией $\omega(x, y) = (x^2 - 1)(y^2 - 1)$ вида (19.68) и зафиксируем две функции из системы функций (19.67):

$$\varphi_1 = x\omega(x, y) = (x^3 - x)(y^2 - 1) \quad \text{и} \quad \varphi_2 = y\omega(x, y) = (x^2 - 1)(y^3 - y).$$

Найдя частные производные

$$\frac{\partial \varphi_1}{\partial x} = (3x^2 - 1)(y^2 - 1), \qquad \frac{\partial^2 \varphi_1}{\partial x^2} = 6x(y^2 - 1),$$

$$\frac{\partial \varphi_1}{\partial y} = 2y(x^3 - x), \qquad \frac{\partial^2 \varphi_1}{\partial y^2} = 2(x^3 - x),$$

$$\frac{\partial \varphi_2}{\partial x} = 2x(y^3 - y), \qquad \frac{\partial^2 \varphi_2}{\partial x^2} = 2(y^3 - y),$$

$$\frac{\partial \varphi_2}{\partial y} = (x^2 - 1)(3y^2 - 1), \qquad \frac{\partial^2 \varphi_2}{\partial y^2} = 6y(x^2 - 1),$$

по формуле (19.70) подсчитываем коэффициенты при неизвестных c_1, c_2 системы типа (19.69). Имеем (с учетом симметрии вхождения переменных x и y в функции φ_1, φ_2 и в их производные, а также симметрии области Ω):

$$a_{11}(= a_{22}) = -\iint_\Omega \varphi_1\left(\frac{\partial^2 \varphi_1}{\partial x^2} + \frac{\partial^2 \varphi_1}{\partial y^2}\right)dxdy =$$

$$= -\int_{-1}^{1}\int_{-1}^{1}(x^3 - x)(y^2 - 1)\left[6x(y^2 - 1) + 2(x^3 - x)\right]dxdy = \frac{3328}{1575},$$

$$a_{12}(= a_{21}) = -\iint_\Omega \varphi_2\left(\frac{\partial^2 \varphi_1}{\partial x^2} + \frac{\partial^2 \varphi_1}{\partial y^2}\right)dxdy =$$

$$= -\int_{-1}^{1}\int_{-1}^{1}(x^2 - 1)(y^3 - y)\left[6x(y^2 - 1) + 2(x^3 - x)\right]dxdy = 0.$$

При нахождении свободных членов b_1, b_2 СЛАУ (19.69) принимаем во внимание, что вывод формул, реализующих метод Ритца, осуществлялся для уравнения Пуассона в записи $-\Delta u = f$ (см. (19.57)), так что здесь следует считать $f(x, y) = x + y$. По формуле (19.71), также учитывая симметрию, получаем

$$b_1(= b_2) = \iint_\Omega \varphi_1 f\, d\Omega = \int_{-1}^{1}\int_{-1}^{1}(x^3 - x)(y^2 - 1)(x + y)dxdy = \frac{16}{45}.$$

Таким образом, система (19.69) в данном случае есть

$$\frac{3328}{1575}c_i = \frac{16}{45}, \quad i = 1, 2,$$

откуда $c_1 = c_2 = \dfrac{35}{208}$. Следовательно, искомое приближенное решение $u_2(x, y)$ в соответствии с формой (19.66) имеет вид

$$u_2(x, y) = \frac{35}{208}x(x^2 - 1)(y^2 - 1) + \frac{35}{208}y(x^2 - 1)(y^2 - 1)$$

или, проще,

$$u_2(x, y) = \frac{35}{208}(x^2 - 1)(y^2 - 1)(x + y). \qquad (19.73)$$

Не пытаясь делать оценок погрешности найденного решения $u_2(x, y)$, вычислим его невязку в одной контрольной точке $(0.5; 0.5)$. Дважды продифференцировав функцию (19.73), получаем функциональное выражение невязки

$$\Delta u_2 + f = \frac{35}{104}\left[(y^2 - 1)(3x + y) + (x^2 - 1)(x + 3y)\right] + (x + y);$$

ее значение в точке $(0.5; 0.5) \in \Omega$ равно $-\dfrac{1}{104}$, что совсем неплохо при таком малом числе координатных функций.

19.7. О ДВУМЕРНОМ МЕТОДЕ КОНЕЧНЫХ ЭЛЕМЕНТОВ

С появлением высокопроизводительной вычислительной техники метод Ритца вряд ли смог бы остаться серьезным конкурентом рассматриваемым в следующих главах конечноразностным методам, если бы не пришло осознание того, что большие области задания уравнений со сложной геометрией можно разбивать на много частей простого вида (***конечные элементы***) и в

качестве координатных функций в методах Ритца и Галёркина использовать базисные сплайны невысокой степени, одномерные, двумерные или трехмерные в зависимости от размерности решаемой задачи, т.е. если бы не пришли к методу, называемому ***методом конечных элементов***. С одномерным МКЭ, применимым к краевым задачам для ОДУ, мы уже познакомились в § 17.6. Сразу же отметим, что двумерный МКЭ, не говоря уже о трехмерном, намного сложнее одномерного, и пакеты компьютерных программ, его реализующие, содержат десятки и сотни тысяч операторов [119]. Более-менее полное изложение этого метода можно найти в книгах [124, 135, 168, 170, 182, 191] и некоторых других. Наша цель — ориентируясь на подготовленные в предыдущем параграфе формулы (19.69)–(19.71), с помощью которых можно получить методом Ритца коэффициенты приближенного решения (19.66) задачи Дирихле для уравнения Пуассона (19.57)–(19.58), лишь попытаться расставить основные вехи на пути освоения метода конечных элементов, что может помочь при дальнейшем его изучении или при использовании готовых программ, его реализующих, если в этом возникнет необходимость.

Метод конечных элементов решения двумерных краевых задач предусматривает несколько этапов.

На первом этапе выполняется разбиение данной области $\Omega \subset \mathbf{R}_2$ на некоторое число простых и однородных по структуре частей Δ_k — конечных элементов. Более редко в этой роли выступают четырехугольники; чаще же конечными элементами Δ_k служат треугольники; в последнем случае, который мы и примем за основу, разбиение области $\overline{\Omega}$ называют ***триангуляцией***. Если $\overline{\Omega}$ — многоугольная область, то триангуляция производится так, чтобы каждые два из получающихся при этом треугольников или не пересекались, или имели общую сторону или общую вершину; при этом не должно быть треугольников, у которых все три вершины лежат на границе $\mathbf{\Gamma}$ области Ω, и не должно быть частей многоугольника, отличных от этих k треугольных элементов (рис. 19.3). Обычно задаются или некоторым числом n строго внутренних (по отношению к $\mathbf{\Gamma}$) вершин P_i $(i = 1, 2, \dots, n)$, общих для нескольких треугольных элементов из множества $\{\Delta_k\}$, что определяет количество координатных функций в приближенном решении (19.66) по методу Ритца, или некоторым числом $h > 0$, ограничивающим максимальный линейный размер треугольника для обеспечения нужной точности. Все эти внутренние вершины P_i (которые определенным образом упорядочиваются) образуют ***сетку узлов*** МКЭ, откуда проистекают его названия ***вариационно-сеточный*** или ***проекционно-сеточный метод***. Если граница $\mathbf{\Gamma}$ области Ω имеет криволинейные участки, то либо в них вписываются ломаные, т.е. предварительно формируется многоугольник, вписанный в $\mathbf{\Gamma}$, либо применяются более сложные алгоритмы МКЭ, допускающие использование криволинейных треугольников.

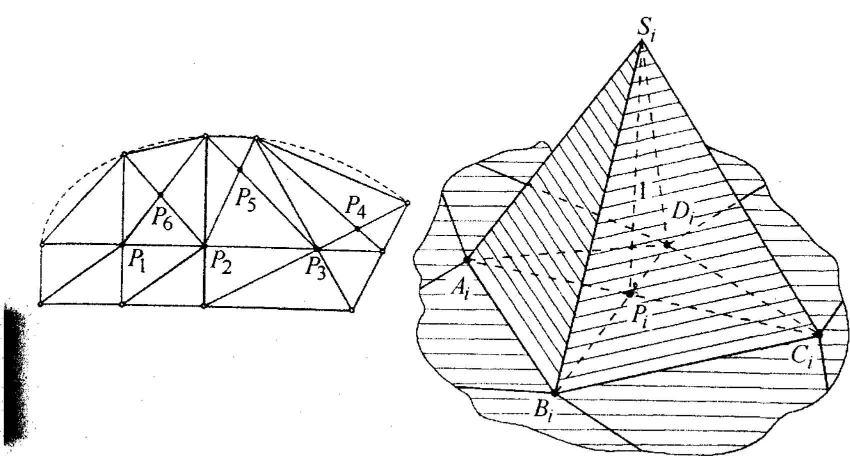


Рис. 19.3. *Пример триангуляции*

Рис. 19.4. *Вид координатной функции φ_i с носителем из четырех лагранжевых треугольных элементов первой степени*

На втором этапе строится система координатных функций φ_i $(i = 1, 2, \dots, n)$, представляющих собой двумерные B-сплайны первой или второй степени, чаще первой. В линейном случае каждую такую функцию φ_i можно изобразить в виде пирамиды с высотой, равной 1 и проведенной из узла сетки P_i, основанием которой, т.е. носителем этой координатной функции, служит многоугольник, состоящий из треугольных элементов с общей вершиной P_i; за пределами этого многоугольника функция φ_i равна нулю (рис. 19.4). Каждая из боковых граней такой пирамиды определена над соответствующим треугольным элементом (одним из $\{\Delta_k\}$) и может быть описана уравнением плоскости, проходящей через ее вершины. Например, грань $A_iB_iS_i$, связанную на рис. 19.4 с элементом $A_iB_iP_i$, можно задать уравнением плоскости

$$z = \alpha_i x + \beta_i y + \gamma_i \tag{19.74}$$

в системе координат $Oxyz$, причем коэффициенты α_i, β_i, γ_i в этом уравнении могут быть однозначно найдены, поскольку первые две координаты у вершин A_i, B_i, S_i известны как координаты узлов A_i, B_i, P_i (какие-то из них могут быть граничными, т.е.

не считаться узлами сетки), а третьи равны 0, 0 и 1 соответственно согласно определению φ_i. Аналогично находятся уравнения плоскостей, соответствующих боковым граням над всеми остальными треугольными элементами с общей вершиной P_i, в результате чего задается кусочно-линейная непрерывная координатная функция φ_i с конечным носителем.

При построении кусочно-квадратичных координатных функций φ_i при той же триангуляции уравнение поверхности над треугольным элементом $A_iB_iP_i$ можно записать в виде

$$z = \alpha_i x^2 + \beta_i xy + \gamma_i y^2 + \delta_i x + \varepsilon_i y + \lambda_i. \tag{19.75}$$

Оно имеет 6 параметров; для их получения используют 6 характерных точек этого элемента, а именно, все его вершины и еще середины сторон. В таком случае тоже однозначно могут быть найдены все коэффициенты в уравнении (19.75) квадратичной поверхности $A_iB_iS_i$ и аналогично остальные уравнения, в комплексе задающие искомую функцию φ_i.

Так как аналитическое построение поверхностей (19.74) и (19.75) естественно считать двумерной лагранжевой интерполяцией, то используемые для этого «трехточечные» и «шеститочечные» треугольные элементы называют ***лагранжевыми элементами*** первой и второй степени соответственно, в отличие, например, от элементов того же геометрического вида, но с информацией о производных в узлах, что отвечает эрмитовой интерполяции*). Треугольные лагранжевы элементы называют также ***треугольниками*** или ***элементами Куранта****) [191].

Линейные (19.74), квадратичные (19.75) и прочие функции двух переменных, определенные показанным выше образом на лагранжевых или иных элементах, называют ***базисными функциями*** соответствующих элементов. Для вычисления коэффициентов базисных функций вместо глобальных координат x, y

---

*) Заметим, что часто под конечным элементом понимается не просто треугольник или четырехугольник, а тройка (иначе, триплет [119]), где первым указывается тип простого элемента, (треугольник или четырехугольник), далее — число используемых в нем узлов (называемое числом степеней свободы) и, наконец, — пространство интерполирующих многочленов (обычно степени ≤ 1 или ≤ 2).

**) Курант Рихард (1888–1972) — математик польского происхождения, профессор Гёттингенского (1920–1933), а затем Нью-Йоркского (с 1934г.) университетов, крупный специалист по краевым задачам для УМФ. Автор ряда книг. Им опубликована первая математическая работа, где используются линейные аппроксимации на тругольных элементах при реализации вариационного метода (1943г.).

обычно применяют локальные так называемые ***барицентрические координаты*** [119, 191].

Н а т р е т ь е м э т а п е применения МКЭ к задаче (19.57)–(19.58) формируется СЛАУ (19.69). При подсчете коэффициентов этой системы по формулам (19.70), (19.71) учитывается выполненное разбиение области Ω на элементарные подобласти (триангуляция), что ввиду аддитивности интегрирования приводит к суммам двойных интегралов по всем отдельным конечным элементам, на каждом из которых определено столько базисных функций, сколько внутренних узлов этот элемент имеет своими вершинами. Так как за пределами определяющего ее элемента базисная функция, а значит, и координатная функция (B-сплайн) за пределами своего носителя — нескольких элементов с общей вершиной, служащей узлом сетки, — равна нулю, то лишь несколько интегралов среди упомянутых будут отличны от нуля. Для их вычисления привлекаются какие-либо простые квадратурные формулы (часто специальные). Ясно, что при большом числе n внутренних вершин элементов, т.е. узлов сетки, матрица $\mathbf{A} := (a_{ij})$ линейной алгебраической системы (19.69) будет разреженной, причем структура этой матрицы (называемой в МКЭ ***матрицей жесткости*** [119, 135, 168, 182]), т.е. расположение в ней ненулевых элементов, зависит от того, каким образом занумерованы узлы и конечные элементы при триангуляции.

Ч е т в е р т ы й э т а п — это выбор подходящего метода и решение полученной СЛАУ. В отличие от одномерного МКЭ, где четко прослеживалась ленточная структура матрицы жесткости (см. § 17.6), здесь на это рассчитывать нельзя, тем более, что двумерный (и трехмерный) МКЭ применяют в случаях областей сложного геометрического вида со сгущением сетки на участках с потенциально быстрым изменением искомого решения. Поэтому часто при реализации МКЭ формирующиеся в нем СЛАУ решаются итерационными методами.

Поставленные выше подзадачи, которые приходится решать в процессе применения МКЭ, обычно рассматриваются в комплексе. Использование специальных приемов разбиения данной области на подобласти с последующим разбиением их на более мелкие конечные элементы позволяет учитывать особенности рассчитываемой модели и готовить для дальнейшей обработки нужные структуры данных [120], а итерационное решение СЛАУ на последовательностях, получающихся при этом сгущающихся сеток делает процесс решения более эффективным, поскольку решения, соответствующие более редким сеткам, можно принимать за начальные приближения для решений на более густых сетках, и, кроме того, процедура сгущения служит основой для апостериорного контроля точности [191].

УПРАЖНЕНИЯ

19.1. А) Завершите построение функции $u_1 = u_1(x, y)$, участвующей в нахождении решения $u = u_1 + u_2$ задачи Дирихле для уравнения Лапласа методом разделения переменных (§ 19.3), считая, что задающие краевые условия (19.23) функции $\varphi_1(x)$ и $\varphi_2(x)$ имеют коэффициенты Фурье в разложении по $\sin\frac{n\pi}{a}x$ соответственно α_n и β_n.

Б) Аналогично построению функции u_1 получите функцию $u_2 = u_2(x, y)$, опираясь на краевые условия (19.24), и запишите окончательный вид решения u задачи (19.21)–(19.22).

19.2. Продолжите пример 19.1 § 19.4 на метод прямых:

а) положив $n = 4$ в первом способе;

б) положив $n = 2$, $x_1 = \frac{1}{3}$, $x_2 = \frac{2}{3}$ во втором способе.

19.3. Запишите расчетные формулы, по которым можно получать сеточные значения приближенных решений системы (19.42) метода прямых для задачи теплопроводности (19.18), (19.19), (19.32), привлекая:

а) явный метод Эйлера;
б) неявный метод Эйлера;
в) метод Рунге–Кутты четвертого порядка.

19.4. Принимая за основу ту или иную из предложенных в § 19.4 концепций метода прямых, рассмотрите возможность его применения к задаче (19.8), (19.16), (19.17) колебания струны [20, 100, 138].

19.5. Конкретизируйте вид функции F, определенной выражением (19.55), в случае $n = 2$; запишите для нее необходимые условия экстремума и убедитесь, что полученная при этом СЛАУ есть система Ритца (19.56) при $n = 2$.

19.6. А) В примере 19.2 § 19.6 пересчитайте коэффициенты a_{11}, a_{22}, пользуясь интегралом Дирихле (19.64); убедитесь в совпадении результатов.

Б) Методом Ритца найдите приближение $u_4(x, y)$ к решению задачи Дирихле примера 19.2, выбрав еще две подходящие координатные функции из системы функций (19.67). Вычислите значение невязки функции $u_4(x, y)$ в точке $(0.5; 0.5)$ и сравните его с имеющимся в примере значением невязки приближения $u_2(x, y)$.

19.7. Считая единственным внутренним узлом сетки точку $P_1(0.5; 0.5)$, сделайте триангуляцию квадрата $[-1, 1] \times [-1, 1]$ и методом конечных элементов найдите приближенное значение $u(0.5, 0.5)$ решения $u(x, y)$ однородной задачи Дирихле для уравнения Пуассона (19.72). Сравните результат со значением $u_2(0.5, 0.5)$, подсчитав последнее по формуле (19.73), полученной в примере 19.2.

ГЛАВА 20 КОНЕЧНОРАЗНОСТНЫЕ МЕТОДЫ РЕШЕНИЯ ЭВОЛЮЦИОННЫХ ЗАДАЧ

С помощью ключевых терминов метода конечных разностей таких, как сетка, шаги сетки, узлы (внутренние и граничные), слой, шаблон разностной схемы и т.п., описываются процессы построения явных и неявных разностных схем для нестационарных задач на основе простейших конечноразностных аппроксимаций производных, изучавшихся в гл. 13. Делается это на примерах одномерного и двумерного уравнений теплопроводности и волнового уравнения; уравнение переноса затрагивается в той мере, в какой это необходимо, чтобы обозначить проблему возникновения разрывных решений и увидеть потребность в консервативных схемах. Особое внимание уделяется конструированию экономичных схем, в частности, методу переменных направлений. Условная или абсолютная устойчивость рассматриваемых разностных схем лишь констатируется (с указанием ссылок на соответствующую литературу).

20.1. НЕКОТОРЫЕ РАЗНОСТНЫЕ СХЕМЫ ДЛЯ УРАВНЕНИЯ ТЕПЛОПРОВОДНОСТИ

Будем рассматривать построение, терминологию и характерные черты ***метода конечных разностей*** для уравнений в частных производных (сначала для нестационарных задач) на примере неоднородного уравнения теплопроводности

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2} + g(x, t), \quad x \in [0, l], \quad t \in [0, T], \qquad (20.1)$$

сопровождаемого начальным по временной переменной t

$$u(x, 0) = \varphi(x) \quad \text{при} \quad x \in [0, l] \qquad (20.2)$$

и краевыми по пространственной переменной x

$$u(0, t) = \alpha(t), \quad u(l, t) = \beta(t) \quad \text{при} \quad t \in [0, T] \qquad (20.3)$$

условиями. Физической интерпретации этой задачи мы касались в предыдущей главе.

Как видим, область Ω, на которой определена данная задача, представляет собой прямоугольник $(0, l) \times (0, T)$ в системе координат Oxt, а ее граница Γ состоит из отрезков прямых $x = 0$, $x = l$, $t = 0$ (отрезок прямой $t = T$ здесь к границе Γ не относится). Разобьем этот прямоугольник на прямоугольные же

части прямыми $x = x_i$ и $t = t_k$, параллельными осям Ot и Ox соответственно, где:

$$x_i = ih, \quad h = \frac{l}{n}, \quad i = 0, 1, \ldots, n; \tag{20.4}$$

$$t_k = k\tau, \quad \tau = \frac{T}{m}, \quad k = 0, 1, \ldots, m. \tag{20.5}$$

Точки $(x_i; t_k) \in \overline{\Omega} := \Omega \cup \boldsymbol{\Gamma}$, лежащие на пересечении этих прямых, называются ***узлами***: ***внутренними***, если они принадлежат области Ω, и ***граничными***, если они лежат на ее границе $\boldsymbol{\Gamma}$ (см. рис. 20.1, где внутренние узлы помечены кружочками, а граничные — крестиками; закрашенные кружочки не относятся ни к внутренним, ни к граничным узлам). Совокупность всех узлов в $\overline{\Omega}$ называют ***сеткой*** для данной задачи (обозначим ее $\overline{\Omega}_h^\tau$), а числа h и τ, фигурирующие в (20.4), (20.5), называют ***шагами сетки*** по переменным x и t соответственно. Узлы, лежащие на одной прямой $t = t_k$ при фиксированном $k = 0, 1, \ldots, m$, называют ***слоем***.

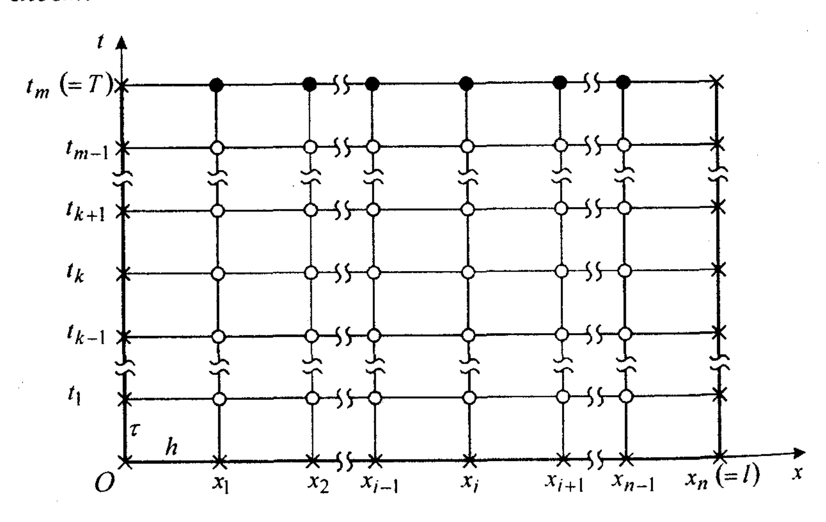


Рис. 20.1. Сетка $\overline{\Omega}_h^\tau$ для конечноразностного метода решения задачи (20.1)–(20.3)

Каждой функции $v(x, t)$, непрерывной в области $\overline{\Omega}$, отвечает единственная таблица ее значений $v_i^k := v(x_i, t_k)$ в узлах сетки, называемая ***сеточной функцией*** (соответствующей функции $v(x, t)$). Пометив номер второй координаты узла у сеточной

функции верхним индексом, означающим номер слоя, таким же образом будем помечать сеточные функции одной временной переменной:

$$\alpha^k := \alpha(t_k), \quad \beta^k := \beta(t_k), \quad k = 0, 1, \ldots, m. \tag{20.6}$$

Разностный метод решения уравнений в частных производных (или, точнее, конечноразностный, поскольку мы используем сетки с постоянным шагом вдоль каждой из осей, что позволяет привлекать конечноразностную интерполяцию) основывается на простой идее построения приближенных сеточных решений: спроектировать данное уравнение на сетку, заменяя входящие в него функции сеточными, а частные производные — их простейшими разностными аппроксимациями. Это проектирование производится следующим образом.

Взяв за основу внутренний узел $(x_i; t_k)$ сетки Ω_h^τ, который будем называть ***расчетным узлом***, поставим в соответствие входящей в уравнение (20.1) функции $g(x, t)$ значение $g_i^k = g(x_i, t_k)$, а частные производные $\frac{\partial^2 u}{\partial x^2}$ и $\frac{\partial u}{\partial t}$ заменим их простейшими разностными аппроксимациями. Ввиду того что здесь частные производные рассматриваются в фиксированной точке, к ним можно применить формулы численного дифференцирования функций одной переменной, изучавшиеся в гл. 13. А именно, для производной $\frac{\partial^2 u}{\partial x^2}$ наиболее естественно употребить симметричную формулу (13.20), согласно которой

$$\left.\frac{\partial^2 u}{\partial x^2}\right|_{\substack{x=x_i \\ t=t_k}} = \frac{u(x_{i-1}, t_k) - 2u(x_i, t_k) + u(x_{i+1}, t_k)}{h^2} + O\left(h^2\right). \tag{20.7}$$

Для приближенной замены $\frac{\partial u}{\partial t}$ априори с одинаковым успехом можно попробовать привлечь простейшие формулы правой (13.15), левой (13.14) и симметричной (13.18) аппроксимации первой производной; соответственно им имеем:

$$\left.\frac{\partial u}{\partial t}\right|_{\substack{x=x_i \\ t=t_k}} = \frac{u(x_i, t_{k+1}) - u(x_i, t_k)}{\tau} + O(\tau), \tag{20.8}$$

$$\left.\frac{\partial u}{\partial t}\right|_{\substack{x=x_i \\ t=t_k}} = \frac{u(x_i, t_k) - u(x_i, t_{k-1})}{\tau} + O(\tau), \tag{20.9}$$

$$\left.\frac{\partial u}{\partial t}\right|_{\substack{x=x_i \\ t=t_k}} = \frac{u(x_i, t_{k+1}) - u(x_i, t_{k-1})}{2\tau} + O\left(\tau^2\right). \tag{20.10}$$

Договоримся о следующем.

Во-первых, будем предполагать, что мы находимся в условиях, когда решение $u(x,t)$ данной задачи существует, единственно и обладает достаточной гладкостью.

Во-вторых, в отличие от оговоренных выше обозначений сеточных функций, будем считать, что

$$u_i^k :\approx u(x_i, t_k), \tag{20.11}$$

т.е. чтобы не вводить другой буквы, через u_i^k обозначаем значения сеточной функции, соответствующей приближенному решению $u_h^\tau(x,t)$ на сетке Ω_h^τ, иначе, ***каркаса*** приближенного, а не точного решения данной задачи.

Подставив в уравнение (20.1) аппроксимации $\frac{\partial^2 u}{\partial x^2}$ по формуле (20.7) и $\frac{\partial u}{\partial t}$ по одной из формул (20.8), (20.9) или (20.10), отбросив погрешности аппроксимаций и учтя при этом обозначение (20.11), для расчетного узла (x_i, t_k) получим следующие уравнения:

$$\frac{u_i^{k+1} - u_i^k}{\tau} = a^2 \frac{u_{i-1}^k - 2u_i^k + u_{i+1}^k}{h^2} + g_i^k, \tag{20.12}$$

$$\frac{u_i^k - u_i^{k-1}}{\tau} = a^2 \frac{u_{i-1}^k - 2u_i^k + u_{i+1}^k}{h^2} + g_i^k, \tag{20.13}$$

$$\frac{u_i^{k+1} - u_i^{k-1}}{2\tau} = a^2 \frac{u_{i-1}^k - 2u_i^k + u_{i+1}^k}{h^2} + g_i^k. \tag{20.14}$$

Будем теперь считать, что расчетный узел $(x_i; t_k)$ смещается по сетке Ω_h^τ, т.е. полагаем в уравнениях (20.12)–(20.14) $i = 1, 2, \ldots, n-1$, $k = 1, 2, \ldots, m-1$. Учитывая, что первыми из используемых в расчетах узлами должны быть точки нулевого слоя, а последними — m-го слоя, дополняем множество возможных здесь значений k значением $k = 0$ для уравнения (20.12) и значением $k = m$ для уравнения (20.13).

Сеточные уравнения, которые получаются в результате аппроксимации производных в данном УМФ разностными отношениями, в совокупности с уравнениями, аппроксимирующими на той же сетке начальные и граничные условия (в данном случае, с учетом обозначений (20.6), это условия*)

$$u_i^0 = \varphi_i \quad (:= \varphi(x_i)), \quad i = 0, 1, \ldots, n, \tag{20.15}$$

$$u_0^k = \alpha^k, \quad u_n^k = \beta^k, \quad k = 0, 1, \ldots, m) \tag{20.16}$$

называются ***разностными схемами***. Конфигурации узлов, в которых связаны одним уравнением разностной схемы значения неизвестной функции — каркаса приближенного решения — называют ***шаблоном разностной схемы***. На рис. 20.2 показаны шаблоны, отвечающие разностным схемам (20.12), (20.13) и (20.14) (а, б и в соответственно).

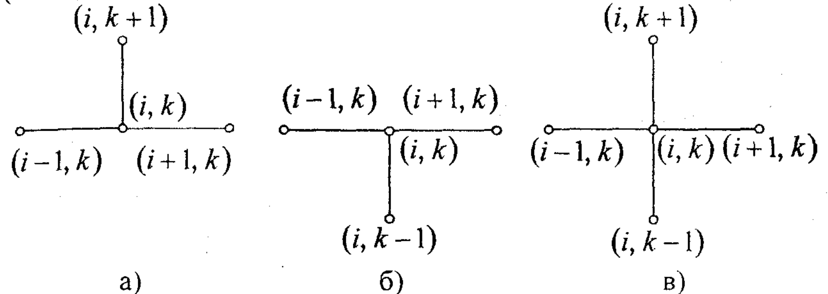


Рис. 20.2. Шаблоны для параболического уравнения: а) явный двухслойный; б) неявный двухслойный; в) явный трехслойный

Введем постоянную (при заданной сетке)

$$\gamma := a^2 \frac{\tau}{h^2} \tag{20.17}$$

и перепишем разностные схемы (20.12)–(20.14) соответственно следующим образом:

$$u_i^{k+1} = \gamma u_{i-1}^k + (1-2\gamma)u_i^k + \gamma u_{i+1}^k + \tau g_i^k$$
$$(\text{где } i = 1, 2, \ldots, n-1; \quad k = 0, 1, \ldots, m-1); \tag{20.18}$$

$$\gamma u_{i-1}^k - (1+2\gamma)u_i^k + \gamma u_{i+1}^k = -u_i^{k-1} - \tau g_i^k$$
$$(\text{где } i = 1, 2, \ldots, n-1; \quad k = 1, 2, \ldots, m); \tag{20.19}$$

$$u_i^{k+1} = u_i^{k-1} + 2\gamma u_{i-1}^k - 4\gamma u_i^k + 2\gamma u_{i+1}^k + 2\tau g_i^k$$
$$(\text{где } i = 1, 2, \ldots, n-1; \quad k = 1, 2, \ldots, m-1). \tag{20.20}$$

---

*) Заметим, что начальное (20.2) и граничные (20.3) условия непременно должны быть согласованными в точках $(0;0)$ и $(l;0)$; так что в (20.15), (20.16)

$$u_0^0 = \alpha^0 = \varphi_0, \quad u_n^0 = \beta^0 = \varphi_n.$$

Эти записи позволяют увидеть, как можно осуществить процесс заполнения $(n-1)\times m$-таблицы значениями u_i^k, определяемыми каждой из трех представленных схем.

Учитывая, что на нулевом слое значения u_i^0, согласно (20.15), известны при любом $i \in \{0, 1, \dots, n\}$, формула (20.18) позволяет непосредственно вычислить все приближенные значения u_i^1 первого слоя*). Таким же образом, привлекая найденные значения u_i^1 первого слоя, непосредственно можно подсчитать значения u_i^2 второго слоя, только при этом для вычислений u_1^2 и u_{n-1}^2 потребуется еще подставить в формулу граничные значения $u_0^1 = \alpha^1$ и $u_n^1 = \beta^1$ соответственно (см. (20.16)). Дальнейшие переходы от слоя к слою не отличаются от этого.

В равенстве (20.19) при $k=1$ и любом $i \in \{1, 2, \dots, n-1\}$ известна только правая часть благодаря начальным данным (20.15). Следовательно, для нахождения значений u_i^1 в узлах первого слоя нужно решить СЛАУ

$$\gamma u_{i-1}^1 - (1+2\gamma)u_i^1 + \gamma u_{i+1}^1 = -\varphi_i - \tau g_i^1,$$
$$\text{где } i = 1, 2, \dots, n-1; \quad u_0^1 = \alpha^1, \quad u_n^1 = \beta^1.$$

Очевидно, матрица этой системы имеет трехдиагональную структуру с диагональным преобладанием $(1+2\gamma > \gamma + \gamma)$, а это означает, что система может быть эффективно решена методом прогонки (§ 2.6). Значения искомой функции u в узлах второго слоя получаются как решение системы

$$\gamma u_{i-1}^2 - (1+2\gamma)u_i^2 + \gamma u_{i+1}^2 = -u_i^1 - \tau g_i^2,$$
$$\text{где} \quad i = 1, 2, \dots, n-1; \quad u_0^2 = \alpha^2, \quad u_n^2 = \beta^2,$$

и так далее, т.е. всего для заполнения таблицы требуется решить m однотипных СЛАУ.

Каждая из двух рассмотренных разностных схем связывает значения искомой функции на двух соседних слоях, откуда проистекает их название ***двухслойные схемы***. При этом разностная схема (20.18), представляющая собой формулу для непосредственного вычисления искомых значений очередного слоя, называется ***явной схемой***, а схема (20.19), требующая при переходе от слоя к слою решать системы алгебраических уравнений, — ***неявной схемой***.

---

*) Эти вычисления можно производить или последовательно, одно за другим, полагая $i = 1, 2, \dots, n-1$, или параллельно, все сразу, при этих же значениях i, что говорит о естественной возможности оптимального использования здесь ресурсов компьютера с параллельной обработкой информации.

Исходя из подобного принципа классификации, *разностная схема* (20.20) *считается* ***явной трехслойной схемой***. Правда, поскольку начать процесс вычислений по этой схеме можно только положив $k=1$, т.е. с формулы

$$u_i^2 = 2\gamma u_{i-1}^1 - 4\gamma u_i^1 + 2\gamma u_{i+1}^1 + \varphi_i + 2\tau g_i^1, \tag{20.21}$$

где значения u_i^1 $(i = 1, 2, \dots, n-1)$ еще не подсчитаны, нужна дополнительная связь между неизвестными и известными величинами. Такую дополнительную связь между значениями искомой сеточной функции трех первых слоев можно получить, например, беря в качестве расчетного узел нулевого слоя и применяя к $\frac{\partial u}{\partial t}$ формулу несимметричной аппроксимации второго порядка точности (13.26). Имеем (представьте соответствующий шаблон!)

$$\frac{-3u_i^0 + 4u_i^1 - u_i^2}{2\tau} = a^2\,\frac{u_{i-1}^0 - 2u_i^0 + u_{i+1}^0}{h^2} + g_i^0,$$

откуда вытекает другое равенство того же вида, что и (20.21):

$$u_i^2 = 4u_i^1 - 2\gamma\varphi_{i-1} + (2\gamma - 3)\varphi_i - 2\gamma\varphi_{i+1} - 2\tau g_i^0.$$

Сравнением правых частей последнего равенства и равенства (20.21) приходим к трехточечному разностному уравнению второго порядка

$$\gamma u_{i-1}^1 - (2+2\gamma)u_i^1 + \gamma u_{i+1}^1 = -\gamma\varphi_{i-1} + (\gamma - 2)\varphi_i - \gamma\varphi_{i+1} - \tau g_i^0 - \tau g_i^1,$$

решая которое методом прогонки, находим значения функции u в узлах первого слоя, после чего становится возможным счет по формуле (20.21) и все последующие вычисления по общей формуле (20.20) при $k = 2, 3, \dots, m-1$.

20.2. АППРОКСИМАЦИЯ, УСТОЙЧИВОСТЬ, СХОДИМОСТЬ РАЗНОСТНЫХ СХЕМ ДЛЯ УРАВНЕНИЯ ТЕПЛОПРОВОДНОСТИ

Формальная аппроксимация производных в уравнении (20.1) на заданной сетке Ω_h^τ, а также дискретизация начальных (20.2) и граничных (20.3) условий в узлах границы $\mathbf{\Gamma}$ области Ω, привели к тому, что данную непрерывную бесконечномерную

задачу (20.1)–(20.3) мы заменили тремя разными конечномерными задачами (20.18)–(20.20) с одинаковыми для них дополнительными условиями (20.15), (20.16). Встает вопрос: какое отношение имеют решения новых задач к решению исходной задачи? Можно ли быть уверенным в том, что последовательность решений u_i^k каждого из сеточных уравнений (20.18)–(20.20) сходится к значениям решения $u(x,t)$ уравнения (20.1) на данной сетке Ω_h^τ при $h \to 0$, $\tau \to 0$? Ответ на поставленные вопросы обычно дают, опираясь на упоминавшееся в § 16.1 утверждение о том, что сходимость есть следствие аппроксимации данной бесконечномерной задачи конечномерной и устойчивости решения последней. При этом доказывается, что быстрота сходимости имеет такой же порядок относительно шага (шагов), что и порядок аппроксимации.

Установление факта и порядка аппроксимации уравнения (20.1), а точнее, задачи (20.1)–(20.3), рассмотренными выше разностными схемами в предположении о достаточной гладкости решения $u(x,t)$ не вызывает затруднений. Сравнение этих схем в виде (20.12)–(20.14) с конечномерными уравнениями, которые получаются в результате подстановки точных равенств (20.7)–(20.10) в исходное уравнение (20.1), показывает, что явная и неявная двухслойные разностные схемы (20.18) и (20.19) аппроксимируют уравнение (20.1) с погрешностью $O(\tau)+O(h^2)$, а явная трехслойная схема (20.20) — с погрешностью $O(\tau^2)+O(h^2)$. Так как при проектировании краевых и начальных условий данной задачи на сетку $\overline{\Omega}_h^\tau$ искажений (зависящих от h и от τ) не вносится, то можно считать, что с такими же погрешностями эта задача в целом аппроксимируется соответствующими дискретными задачами вместе с условиями (20.15)–(20.16). Как правило, о рассмотренных двухслойных схемах, основанных на четырехточечных шаблонах (см. рис. 20.2, а и б), говорят, что они аппроксимируют данную параболическую задачу (20.1)–(20.3) со вторым порядком по пространственной переменной x и с первым по временной переменной t, и отражают этот факт записью погрешности аппроксимации задачи вида $O(h^2+\tau)$. Для трехслойной схемы (20.20) погрешность аппроксимации составляет величину $O(h^2+\tau^2)$.

Несомненно, погрешность аппроксимации данной задачи с помощью каждой из трех рассматриваемых разностных схем (в

другой терминологии, ***локальная ошибка дискретизации*** [138]) стремится к нулю при $h \to 0$, $\tau \to 0$. Это называется ***условием согласованности разностной схемы*** и является лишь необходимым условием стремления к нулю глобальной ошибки, которая может накапливаться по тому или иному закону от слоя к слою. Способность разностной схемы (естественно, при условии ее однозначной разрешимости) не допускать неограниченного увеличения ошибки в процессе измельчения сетки, грубо говоря, и означает ее устойчивость. Поскольку источником первоначальной ошибки может служить переход к сеточным функциям при дискретизации исходного уравнения, а также неточности в начальных и граничных условиях, изучение устойчивости разностной схемы в целом разбивают, соответственно, на изучение устойчивости по правой части, по начальным данным и по граничным условиям.

Исследование устойчивости разностных схем (в смысле той или иной строгой формулировки, которых имеется несколько, но все они базируются на одном и том же понятии устойчивости) является наиболее сложным этапом построения конечноразностного метода. Существует ряд приемов проведения таких исследований [20]: с помощью разностного принципа максимума, с помощью «индекса разностной схемы», методом разделения переменных, изучением роста единичной ошибки и др. Сошлемся на распространенную литературу, где можно найти изучение устойчивости предложенных выше схем разными приемами и в разных подробностях : в [13, 14, 20, 103, 154, 158, 159] — более подробное, на наш взгляд, изложение, в [27, 44, 47, 62, 78, 92, 100, 126, 138, 153, 178] — менее подробное.

Обратимся к последней разностной схеме (20.20). К ней целесообразно применить так называемую ***ε-схему*** изучения роста единичной ошибки [20, 27, 84]. Суть ее в том, что по исследуемой разностной схеме проводят пробные расчеты, исходя из предположения о единственной ошибке ε в значении u_i^k, т.е. вместо этого значения в расчетную формулу подставляют значение $u_i^k+\varepsilon$ и следят за поведением ошибки на следующих $(k+1)$-м, $(k+2)$-м, ... слоях. Если ошибка имеет тенденцию расти по модулю, то разностная схема должна быть признана неустойчивой и забракована. Такой ***ε-анализ*** трехслойной схемы (20.20) в частном случае $\gamma=0.5$, $g\equiv 0$ показывает [20], что начав процесс вычислений с k-го слоя с одним искаженным значением $u_i^k+\varepsilon$, на следующем слое

на следующем слое получим значение u_i^{k+1} с ошибкой -2ε, далее значение u_i^{k+2} будет иметь ошибку 7ε, ..., значение u_i^{k+6} — ошибку 1311ε. Ошибка катастрофически растет (и расползается по таблице подобно тому, как это наблюдалось с поведением ошибки в таблице конечных разностей, см. табл. 8.6. в § 8.4), что говорит о непригодности разностной схемы (20.20), по меньшей мере, при $\gamma = 0.5$, хотя она имеет более высокий порядок аппроксимации по сравнению с двумя другими схемами.

Исследования явной двухслойной схемы (20.18) приводят к одному: она устойчива при условии $\gamma \le 0.5$, что в соответствии с обозначением (20.17) равносильно требованию к шагу по времени

$$\tau \le \frac{h^2}{2a^2}. \qquad (20.22)$$

Разностная схема, устойчивость которой связана с некоторым ограничением на шаг, называется ***условно устойчивой***.

В отличие от явной, неявная двухслойная схема (20.19) устойчива при любых $\gamma > 0$, т.е. при любом соотношении шагов по времени и по пространственной переменной, в связи с чем ее называют ***безусловно*** или ***абсолютно устойчивой разностной схемой***.

Таким образом, можно констатировать сходимость решений двухслойных разностных схем (20.18), (20.19) с совокупностью дополнительных условий (20.15)–(20.16) к решению задачи теплопроводности (20.1)–(20.3) с погрешностью $O(h^2 + \tau)$ при любых $h \to 0$, $\tau \to 0$ в случае неявной схемы и при $h \to 0$, $\tau \le \frac{h^2}{2a^2} \to 0$ в случае явной схемы. Очевидно, каждая из этих схем имеет свои достоинства и недостатки. Явная схема проще и требует меньше вычислительных затрат на подсчет значений одного слоя, зато таких слоев должно быть больше из-за ограничения (20.22) на шаг по времени, чем при реализации неявной схемы, допускающей любое соотношение шагов, но требующей решения СЛАУ при подсчете значений каждого слоя. Правда, при этом сопоставлении не следует забывать еще о точности аппроксимации данной задачи разностной схемой, которая для явной схемы (20.18) при ограничении (20.22) теперь есть $O(h^2)$, и чтобы иметь такую же для неявной схемы, нужно в ней брать $\tau := O(h^2)$ [47].

20.3. ДВУХСЛОЙНЫЙ ШЕСТИТОЧЕЧНЫЙ И ДРУГИЕ ШАБЛОНЫ ДЛЯ ПАРАБОЛИЧЕСКИХ УРАВНЕНИЙ

Возьмем за основу следующего построения явную и неявную разностные схемы для уравнения (20.1) в их исходной форме (20.12) и (20.13). Перепишем равенство (20.13) в виде

$$\frac{u_i^{k+1} - u_i^k}{\tau} = a^2 \frac{u_{i-1}^{k+1} - 2u_i^{k+1} + u_{i+1}^{k+1}}{h^2} + g_i^{k+1} \qquad (20.23)$$

(т.е. увеличим в нем на единицу верхний индекс) и сравним результат с сеточным уравнением (20.12). Видим, что в левых частях (20.12) и (20.23) стоит одна и та же аппроксимация производной $\frac{\partial u}{\partial t}$, пригодная для всего отрезка $[t_k, t_{k+1}]$ прямой $x = x_i$ и имеющая наивысшую точность $O(\tau^2)$ в средней точке $t_k + 0.5\tau$ этого отрезка (см. § 13.2). Дроби в правых частях этих уравнений представляют собой аппроксимации второй производной $\frac{\partial^2 u}{\partial x^2}$ одного типа, но на разных слоях: на слое k в уравнении (20.12) и на слое $k+1$ в уравнении (20.23); иначе, расчетными точками служат точки $(x_i; t_k)$в первом и $(x_i; t_{k+1})$ во втором случаях. Есть некий смысл в том, чтобы в качестве расчетной точки при построении новой схемы взять не узловую точку прямой $x = x_i$, а какую-то промежуточную точку отрезка $[t_k, t_{k+1}]$ этой прямой. В зависимости от того, в каком отношении эта условная расчетная точка будет делить указанный отрезок, соединяющий k-й и $(k+1)$-й слои, производную $\frac{\partial^2 u}{\partial x^2}$ будем подменять соответствующей линейной комбинацией ее аппроксимаций на слоях k и $k+1$; ту же линейную комбинацию применяем к сеточной функции $g(x_i, t_k)$. Таким образом, приходим к равенству*)

$$\frac{u_i^{k+1} - u_i^k}{\tau} = \frac{a^2}{h^2}\Big[\sigma\left(u_{i-1}^k - 2u_i^k + u_{i+1}^k\right) +$$
$$+ (1-\sigma)\left(u_{i-1}^{k+1} - 2u_i^{k+1} + u_{i+1}^{k+1}\right)\Big] + \sigma g_i^k + (1-\sigma) g_i^{k+1}, \qquad (20.24)$$

---

*) По логике предшествующих рассуждений относительно условной расчетной точки вместо двух последних слагаемых в нем лучше использовать значение $g(x_i, \sigma t_k + (1-\sigma)t_{k+1})$.

где $\sigma \in [0, 1]$ — вещественный параметр (*вес*); $i = 1, 2, \ldots, n-1$; $k = 0, 1, \ldots, m-1$.

Ясно, что вновь построенная разностная схема (20.24), называемая ***схемой с весами***, обобщает схемы (20.12) и (20.23): при $\sigma = 1$ — это явная схема (20.12), при $\sigma = 0$ — неявная схема (20.23) или, что то же, (20.13). Если $0 < \sigma < 1$, то равенство (20.24) связывает шесть точек двух слоев, т.е. отвечает шаблону, изображенному на рис. 20.3.

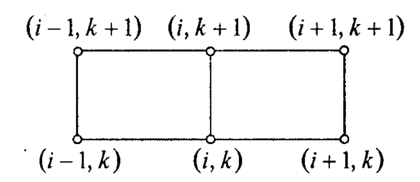


Рис. 20.3. *Двухслойный шеститочечный шаблон для разностной схемы* (20.24)

Заметим, что явной схема (20.24) является только при значении веса $\sigma = 1$; при всех остальных значениях $\sigma \in [0, 1]$ она — неявная. Чтобы отличить от всех неявных схем этого однопараметрического семейства рассмотренную ранее неявную схему (20.13), соответствующую четырехточечному шаблону, ее зачастую называют ***чисто неявной схемой***. Для неявных схем применяют также названия ***схемы с опережением*** [158] или ***схемы с упреждением*** [172].

Другим важным частным случаем семейства формул (20.24) является случай $\sigma = 0.5$. При этом значении σ с учетом обозначения постоянной γ (20.17) из (20.24) имеем равенство

$$u_i^{k+1} - u_i^k = \frac{\gamma}{2}\left(u_{i-1}^k - 2u_i^k + u_{i+1}^k\right) + \\ + \frac{\gamma}{2}\left(u_{i-1}^{k+1} - 2u_i^{k+1} + u_{i+1}^{k+1}\right) + \frac{\tau}{2}\left(g_i^k + g_i^{k+1}\right),$$

которому далее придаем типичный вид трехточечного разностного уравнения второго порядка относительно неизвестных значений на $(k+1)$-м слое:

$$\gamma u_{i-1}^{k+1} - (2 + 2\gamma)u_i^{k+1} + \gamma u_{i+1}^{k+1} = G_i^k, \qquad (20.25)$$

где

$$G_i^k := -\gamma u_{i-1}^k - (2 - 2\gamma)u_i^k - \gamma u_{i+1}^k - \tau g_i^k - \tau g_i^{k+1}. \qquad (20.26)$$

Полученная неявная двухслойная разностная схема (20.25)–(20.26) (в которой полагаем $i = 1, 2, \ldots, n-1$; $k = 0, 1, \ldots, m-1$) называется ***схемой Кранка–Николсон****). Практически очевидно, что она аппроксимирует уравнение (20.1) с точностью $O\left(h^2 + \tau^2\right)$, а технология ее использования не отличается от описанной выше для чисто неявной схемы (20.19). Доказана абсолютная устойчивость этой схемы, что позволяет проводить расчеты по ней с произвольными шагами h и τ, обеспечивающими достаточную точность аппроксимации.

При любых других значениях $\sigma \in [0, 1]$, $\sigma \neq 0.5$, погрешность аппроксимации схемы с весами (20.24), (называемой также ***обобщенной схемой Кранка–Николсон***), составляет величину $O\left(h^2 + \tau\right)$. Для $\sigma \in [0, 0.5]$ схема абсолютно устойчива, если же $\sigma \in (0.5, 1]$, то устойчивость имеет место при выборе шага по времени, удовлетворяющего неравенству $\tau \le \dfrac{h^2}{2a^2(2\sigma - 1)}$ (доказательство см., например, в [13]). Схема Кранка–Николсон выделяется из семейства (20.24) как самая предпочтительная [10].

Наряду с рассмотренными выше применяют и другие схемы аппроксимации уравнения теплопроводности (20.1) с менее прозрачной логикой построения. Вот две из них [100].

1. ***Неявная пятиточечная трехслойная схема***

$$\frac{3}{2}\frac{u_i^{k+1} - u_i^k}{\tau} - \frac{1}{2}\frac{u_i^k - u_i^{k-1}}{\tau} = a^2\frac{u_{i-1}^{k+1} - 2u_i^{k+1} + u_{i+1}^{k+1}}{h^2} + g_i^{k+1}; \qquad (20.27)$$

абсолютно устойчива, погрешность аппроксимации $O\left(h^2 + \tau^2\right)$.

2. ***Схема Дюфорта и Франкела***

$$\frac{u_i^{k+1} - u_i^{k-1}}{2\tau} = a^2\frac{u_{i+1}^k - u_i^{k+1} - u_i^{k-1} + u_{i-1}^k}{h^2} + g_i^k \qquad (20.28)$$

— явная четырехточечная трехслойная; сходится с порядком $O\left(h^2 + \tau\right) + O\left(\dfrac{\tau^2}{h^2}\right)$ при условии, что $\tau \to 0$, $h \to 0$, $\dfrac{\tau}{h} \to 0$.

---

*) Авторы статьи [202], в которой впервые рассмотрена данная схема, английские математик John Crank (род. 1916) и физик Phyllis Nicolson (1917–1968). Учитывая, что второй из авторов этой публикации — женщина, обращаем внимание на бытующее в отечественной литературе грамматически неправильное написание «схема Кранка–Николсона».

Шаблоны этих схем изображены на рис. 20.4, а и б соответственно.

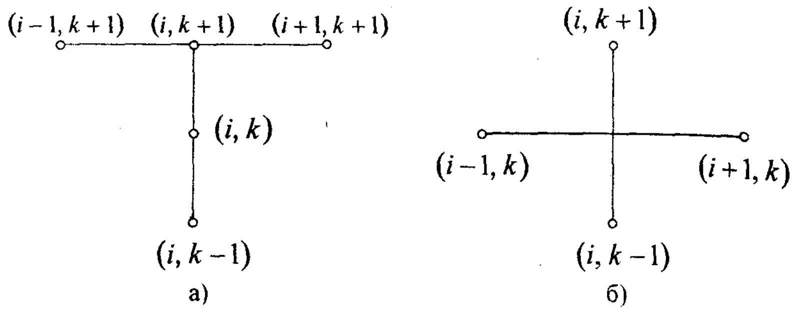


Рис. 20.4. *Шаблоны разностных схем* (20.27) *и* (20.28)

20.4. ДИСКРЕТИЗАЦИЯ ВОЛНОВОГО УРАВНЕНИЯ

Рассмотрим применение метода конечных разностей к задаче о свободных колебаниях струны:

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}, \tag{20.29}$$

$$u(x, 0) = \varphi(x) \quad \text{при } x \in [0, l], \tag{20.30}$$

$$\frac{\partial u(x, 0)}{\partial t} = \psi(x) \quad \text{при } x \in [0, l], \tag{20.31}$$

$$u(0, t) = 0, \quad u(l, t) = 0 \quad \text{при } t \in [0, T]. \tag{20.32}$$

О смысле каждого из равенств (20.29)–(20.32) говорилось в § 19.2. Ясно, что начальное условие (20.30) и граничные условия (20.32) должны быть согласованы в точках $(0; 0)$ и $(l; 0)$.

Данная задача определена в той же прямоугольной области $\Omega := (0, l) \times (0, T)$ с границей $\mathbf{\Gamma}$, что и задача теплопроводности (20.1)–(20.3), и для построения разностных схем здесь естественно использовать те же сетки Ω_h^τ внутренних узлов $(x_i; t_k)$ и $\overline{\Omega}_h^\tau$ всех узлов области $\overline{\Omega} := \Omega \cup \mathbf{\Gamma}$, где x_i, t_k задаются равенствами (20.4), (20.5). Так же естественно аппроксимировать $\frac{\partial^2 u}{\partial t^2}$ по формуле вида (20.7), посредством которой аппроксимируется $\frac{\partial^2 u}{\partial x^2}$, т.е. по формуле

$$\left.\frac{\partial^2 u}{\partial t^2}\right|_{\substack{x=x_i \\ t=t_k}} = \frac{u(x_i, t_{k-1}) - 2u(x_i, t_k) + u(x_i, t_{k+1})}{\tau^2} + O\left(\tau^2\right).$$

В таком случае, имея в виду прежнее обозначение (20.11)

$$u_i^k :\approx u(x_i, t_k),$$

приходим к разностной схеме

$$\frac{u_i^{k-1} - 2u_i^k + u_i^{k+1}}{\tau^2} = a^2 \frac{u_{i-1}^k - 2u_i^k + u_{i+1}^k}{h^2},$$

аппроксимирующей волновое уравнение (20.29) на шаблоне типа «крест» (рис. 20.2в) с точностью $O\left(h^2 + \tau^2\right)$. Обозначив

$$\widetilde{\gamma} := a^2 \frac{\tau^2}{h^2}, \tag{20.33}$$

эту схему преобразуем к простой явной формуле

$$u_i^{k+1} = -u_i^{k-1} + \widetilde{\gamma} u_{i-1}^k + (2 - 2\widetilde{\gamma}) u_i^k + \widetilde{\gamma} u_{i+1}^k, \tag{20.34}$$

где полагаем $i = 1, 2, \ldots, n-1$; $k = 1, 2, \ldots, m-1$.

Чтобы однозначно осуществить процесс вычислений по формуле (20.34) при фиксированных n и m (или h и τ), т.е. получить таблицу значений u_i^k — приближенный каркас решения $u(x_i, t_k)$ данной задачи на сетке Ω_h^τ, нужно дополнить эту формулу значениями на нулевом слое

$$u_0^0 = 0, \quad u_1^0 = \varphi_1, \ldots, \quad u_{n-1}^0 = \varphi_{n-1}, \quad u_n^0 = 0 \tag{20.35}$$

и значениями

$$u_0^k = 0, \quad u_n^k = 0 \quad (k = 1, 2, \ldots, m) \tag{20.36}$$

на границе (что соответствует заданным условиям (20.30) и (20.32)), а также воспользоваться какой-нибудь аппроксимацией производной в условии (20.31) для подсчета значений u_i^1 на первом слое. Самое простое — применить здесь простейшую несимметричную аппроксимацию

$$\frac{\partial u(x_i, 0)}{\partial t} = \frac{u(x_i, t_1) - u(x_i, 0)}{\tau} + O(\tau);$$

тогда, согласно (20.31), имеем

$$\frac{u_i^1 - u_i^0}{\tau} = \psi_i \quad \Rightarrow \quad u_i^1 = u_i^0 + \tau \psi_i, \tag{20.37}$$

и далее можно вести счет по формуле (20.34), привлекая по ходу вычислений равенства (20.35), (20.36). Но при таком способе

подсчета значений на первом слое точность аппроксимации данной задачи разностной схемой (20.34)–(20.37) в целом будет лишь $O(h^2 + \tau)$, поскольку «на старте» качество аппроксимации оказывается более низким.

Вместо (20.37) можно получить формулу второго порядка точности. Выведем ее ***методом фиктивной точки***, суть которого — в следующем.

При каждом $i = 1, 2, \ldots, n-1$ будем использовать точку $(x_i; t_{-1}) \equiv (x_i; -\tau)$, лежащую за пределами сеточной области $\overline{\Omega}_h^\tau$, и привлекать какие-либо два соотношения, опирающиеся на данные в задаче условия и связывающие значение $u_i^{-1} :\approx u(x_i, t_{-1})$ искомой функции в этой «фиктивной точке» со значениями u_i^0 и u_i^1 на нулевом и первом слоях; исключение этого фиктивного значения u_i^{-1} приводит к дополнительному уравнению относительно значений u_i^0 и u_i^1.

В данном случае это можно реализовать так.

Пользуясь тем, что по формуле симметричной аппроксимации

$$\left.\frac{\partial u}{\partial t}\right|_{\substack{x=x_i\\ t=0}} = \frac{u(x_i, t_1) - u(x_i, t_{-1})}{2\tau} + O(\tau^2),$$

в силу условия (20.31) можно записать равенство

$$\frac{u_i^1 - u_i^{-1}}{2\tau} = \psi_i,$$

откуда

$$u_i^{-1} = u_i^1 - 2\tau\psi_i. \tag{20.38}$$

Второе из требуемых соотношений получаем из основного равенства (20.34), полагая в нем $k = 0$ и учитывая сеточные начальные данные (20.35). Имеем:

$$u_i^1 = -u_i^{-1} + \tilde{\gamma} u_{i-1}^0 + (2 - 2\tilde{\gamma}) u_i^0 + \tilde{\gamma} u_{i+1}^0 \quad \Rightarrow$$

$$u_i^{-1} = -u_i^1 + \tilde{\gamma}\varphi_{i-1} + (2 - 2\tilde{\gamma})\varphi_i + \tilde{\gamma}\varphi_{i+1}. \tag{20.39}$$

Приравнивая правые части в выражениях фиктивного значения u_i^{-1} по формулам (20.38) и (20.39), приходим к явной формуле второго порядка точности для вычисления недостающих для

применения схемы (20.34) значений на первом слое:

$$u_i^1 = \tau\psi_i + (1 - \tilde{\gamma})\varphi_i + 0.5\tilde{\gamma}(\varphi_{i-1} + \varphi_{i+1}). \tag{20.40}$$

Последнее равенство можно получить и другим путем, например, разлагая функцию $u(x_i, t)$ в точке $(x_i; 0)$ по степеням t и используя разностную аппроксимацию $\dfrac{\partial^2 u(x_i, 0)}{\partial x^2}$ вместо $\dfrac{\partial^2 u(x_i, 0)}{\partial t^2}$ ввиду их исходной связи (20.29) [158].

Доказано [78, 100, 103, 138, 158] (как правило, это делается методом разделения переменных для разностной задачи), что *условие*

$$\tilde{\gamma} \le 1, \quad \textit{т.е.} \quad \tau \le \frac{h}{a}, \tag{20.41}$$

обеспечивает устойчивость построенной явной трехслойной разностной схемы *). Как видим, здесь опять фигурирует ограничение на величину шага по времени, но уже не такое жесткое, как в случае явных схем для параболических уравнений (шаг по времени может быть того же порядка, что и шаг по пространственной переменной). Таким образом, при условии (20.41) имеет место сходимость

$$\left\| u_i^k - u(x_i, t_k) \right\| \to 0 \quad \text{при} \quad h \to 0, \quad \tau \to 0$$

решений сеточных уравнений (20.34) (с $\tilde{\gamma}$ из (20.33)) с дополнительными равенствами (20.35), (20.36); если при этом используется равенство (20.37), то погрешность каркаса приближенного решения составляет величину $O(h^2 + \tau)$, если (20.40) — $O(h^2 + \tau^2)$.

Безусловно устойчивые разностные схемы для гиперболических уравнений существуют, но все они являются неявными. Здесь у них нет больших преимуществ перед явными, и они применяются крайне редко [138].

---

*) Условие $a\tau < h$ называют ***условием Куранта***. В случае $a\tau < h$ отмечается «слабая неустойчивость счета» по явной формуле (20.34) [78].

20.5. О КОНСЕРВАТИВНЫХ СХЕМАХ И О РАЗРЫВНЫХ РЕШЕНИЯХ

Рассмотренные выше построения разностных схем для двух наиболее простых и хорошо изученных двумерных нестационарных задач математической физики на основе формальной замены производных разностными отношениями могут служить лишь некоторым образцом, отправной точкой при приближенном решении конечноразностным методом тех или иных конкретных содержательных задач. Более полное представление о проблемах, возникающих на этом пути, и о способах их разрешения можно получить, изучая более подробную в этой части учебную и специальную научную литературу (см., например, [9, 14, 20, 27, 78, 102, 113, 143, 156, 159, 160]). Здесь мы попытаемся только вкратце ознакомиться с понятием консервативной схемы и с одним из источников появления разрывных решений, что желательно иметь в виду при конструировании разностных схем для нестационарных задач.

1. О консервативных схемах [92, 152, 178]. Если требующее решения дифференциальное уравнение с частными производными является не абсолютной абстракцией, а служит математической моделью некоторого физического явления, то, надо понимать, оно может иметь обычную форму записи относительно искомой функции (плотности, скорости, давления, температуры и т.п.), а может быть записано в так называемом ***дивергентном виде***, где искомым параметром выступает масса, импульс, энергия, ... Дивергентная запись уравнения отражает в дифференциальной форме соответствующий локальный (т.е. в текущей точке пространства в текущий момент времени) физический закон сохранения величин и может быть выведена из глобального (т.е. в некоторой конечной области пространства в некоторый конечный промежуток времени) закона сохранения, имеющего интегральную форму записи. Учитывая более общий характер глобальных законов сохранения, целесообразно конструировать разностные схемы для УМФ так, чтобы они в ячейках сетки удовлетворяли соответствующим интегральным соотношениям и при суммировании по всем ячейкам давали закон сохранения по всей области в целом. Такие разностные схемы называют ***консервативными*** (или, реже, ***дивергентными***). Построение консервативных схем более сложно и требует применения специальных приемов. Одним из них является ***интегро-интерполяционный метод*** (***метод баланса***), основанный на записи интегральных уравнений, отражающих закон сохранения для элементарных ячеек, и применении к интегралам простейших квадратурных формул [158, 159] (нечто подобное мы уже делали ранее в гл. 14, 15 при построении численных процессов решения обыкновенных дифференциальных уравнений, правда, сугубо на формальной основе).

Консервативные разностные схемы более точно передают свойства искомого решения, заложенные в соответствующих законах сохранения, чем неконсервативные схемы. Особо это важно в случае, когда искомое решение может быть разрывным, что естественно допускается интегральной формулировкой закона и лишь формально описывается дифференциальным уравнением. В такой ситуации аппроксимирующая уравнение устойчивая неконсервативная схема, имеющая при фиксированных шагах единственное решение, может приводить в пределе к функциям, не имеющим ничего общего с искомым обобщенным решением (есть сходимость, но не к тому), в то время как на сходимость к этому разрывному решению консервативной схемы, удовлетворяющей условиям аппроксимации и устойчивости, можно рассчитывать наверняка. Для линейных УМФ с постоянными коэффициентами и гладкими решениями консервативность схем, получаемых конечноразностной аппроксимацией производных, обычно имеет место.

Кроме консервативности, имеется несколько других свойств разностных схем, которые связывают с соответствующими свойствами решений и которые полезно учитывать в конкретных задачах: монотонность, положительность и т.п.

2. О характеристиках и разрывных решениях. Затронутая выше проблема нахождения разрывных решений особо актуальна и показательна для квазилинейных уравнений переноса (т.е. уравнений вида (19.12), где под c понимается не постоянная, а некоторая функция, зависящая от искомой переменной u), а также для уравнений и систем гиперболического типа.

Сначала рассмотрим на полуплоскости

$$-\infty < x < \infty, \quad t \geq 0$$

линейное однородное уравнение переноса

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0 \qquad (20.42)$$

с начальным условием

$$u(x, 0) = \varphi(x). \qquad (20.43)$$

В данной полуплоскости выделим те линии $x = x(t)$, которые удовлетворяют обыкновенному дифференциальному уравнению

$$\frac{dx}{dt} = c.$$

Очевидно, этому уравнению удовлетворяют всевозможные прямые

$$x = ct + d, \qquad (20.44)$$

где $d \in (-\infty, +\infty)$ — произвольная постоянная. Прямые (20.44) называются ***характеристиками*** уравнения (20.42). Придавая

параметру d различные вещественные значения, можно получать сколь угодно густую сеть параллельных прямых: наклоненных под острым углом по отношению к положительному направлению оси Ox, если $c > 0$, и под тупым, если $c < 0$ (рис. 20.5).

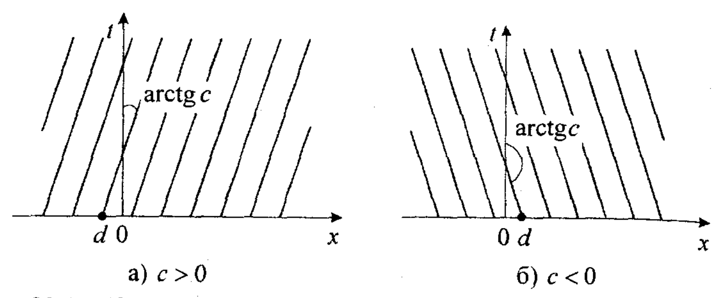


Рис. 20.5. Характеристики линейного уравнения переноса (20.42)

Предположим, что функция $u(x,t)$, служащая решением уравнения (20.42), рассматривается на характеристике (20.44). Тогда, в силу связи $x = x(t)$, она будет функцией только одной переменной t:

$$u(x, t) = u(x(t), t) =: \tilde{u}(t).$$

Ее дифференцирование по t с привлечением формулы полной производной дает

$$\frac{d\tilde{u}}{dt} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} \cdot \frac{dx}{dt} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} c,$$

что с учетом удовлетворения уравнению (20.42) означает

$$\frac{d\tilde{u}}{dt} = 0 \quad \Rightarrow \quad \tilde{u}(t) \equiv const.$$

Таким образом, приходим к выводу, что решение уравнения (20.42) на его характеристиках сохраняет постоянные значения. Другими словами, начальные значения, задаваемые условием (20.43) в момент времени $t = 0$, при увеличении t переносятся вдоль характеристик, и если начальная функция $\varphi(x)$ разрывна, то, значит, и решение $u(x, t)$ будет иметь разрыв вдоль характеристик, проходящих через точки разрыва $\varphi(x)$, что должно учитываться при построении разностных схем для задачи (20.42)-(20.43).

Замечание 20.1. Проведенные выше рассуждения показывают, что при постановке начально-граничной задачи для уравнения переноса (20.42) (или более общего (19.12)) в прямоугольнике $[0, l] \times [0, T]$ краевое условие на функцию $u(x, t)$ достаточно задавать только на одной стороне прямоугольника: при $x = 0$, если $c > 0$, и при $x = l$, если $c < 0$.

Для квазилинейного уравнения переноса

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0 \tag{20.45}$$

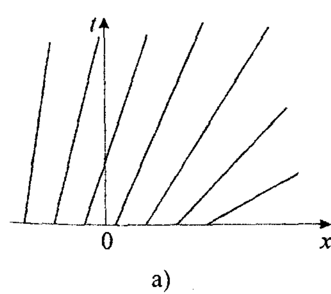
характеристики задаются с помощью ОДУ $\frac{dx}{dt} = u$, и при каждом фиксированном значении u представляют собой семейства прямых

$$x = ut + d \tag{20.46}$$

с параметром семейства d. Эти характеристики обладают тем же свойством, что и прямые (20.44), передавать неизменным значение решения от меньших значений временной переменной к большим значениям. Придавая параметру d те или иные вещественные значения, будем иметь при заданном начальном условии (20.43) конкретные значения

$$u = u(d, 0) = \varphi(d)$$

и тем самым фиксировать конкретные прямые из семейства (20.46). Поскольку их угловые коэффициенты $u = \varphi(d)$ изменяются с изменением d, квазилинейное уравнение (20.45) имеет уже не параллельные характеристические прямые, как в случае линейного уравнения (20.42), а веерообразные, расходящиеся с ростом t, если начальная функция $\varphi(x)$ — возрастающая, и сближающиеся с ростом t, если $\varphi(x)$ — убывающая (рис. 20.6).



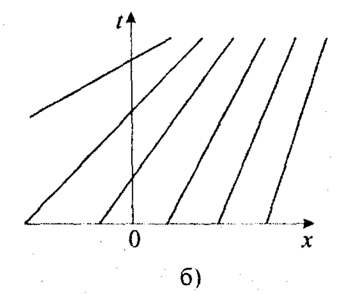


Рис. 20.6. Характеристики квазилинейного уравнения переноса (20.45):
а) случай возрастающей $\varphi(x)$ *в* (20.43);
б) случай убывающей $\varphi(x)$

Если ситуация, когда функция $\varphi(x)$ возрастает, малоприме-

чательна, то при убывании $\varphi(x)$ неизбежно появление разрывных решений (хотя $\varphi(x)$ непрерывна). Действительно, как видно из рис. 20.6б, любые две характеристики уравнения (20.45) обязательно пересекутся, и в точку пересечения эти характеристики перенесут с оси Ox каждая свое значение решения. Возникающее противоречие в том, что считать решением данной задачи на линиях, образованных точками пересечения характеристик (линиях разрыва), требует, во-первых, доопределения самого понятия решения в таких точках (здесь обычно привлекаются интегральные законы сохранения), а во-вторых, учета наличия разрывных решений при конструировании разностных схем. Один из способов такого учета при составлении разностных уравнений, например, для систем уравнений с частными производными гиперболического типа, имеющих два семейства характеристик, связан с построением специальных сеток, узлами которых служат точки пересечения характеристик (в общем случае криволинейных) или согласованные с ними определенным образом точки. Первичные сведения об этом методе, называемом ***методом характеристик***, можно найти в учебных пособиях [152, 178], подробное изложение и библиографию — в монографии [113].

20.6. РАЗНОСТНЫЕ СХЕМЫ ДЛЯ ПАРАБОЛИЧЕСКОГО УРАВНЕНИЯ С ДВУМЯ ПРОСТРАНСТВЕННЫМИ ПЕРЕМЕННЫМИ

Посмотрим, что означает для метода конечных разностей повышение на единицу размерности решаемой параболической задачи по сравнению с изучавшейся в § 20.1 аналогичной задачей с одной пространственной переменной. Сделаем это на примере задачи теплопроводности для изотропной*) квадратной пластинки $\Omega := [0,1]\times[0,1]$ (границу которой будем обозначать $\boldsymbol{\Gamma}_\Omega$) в координатной плоскости Oxy для промежутка времени $t\in[0,T]$. Примем за данность, что процесс теплопередачи описывается уравнением

$$\frac{\partial u}{\partial t} = a^2\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right)+f(x,y,t), \qquad (20.47)$$

*) Неизотропный случай описан, например, в [92].

определенным в параллелепипеде $\Omega\times[0,T]$ (рис. 20.7), начальным распределением температуры

$$u(x,y,0)=\varphi(x,y) \quad \text{при} \quad (x;y)\in\Omega \qquad (20.48)$$

и значениями температуры на границе пластинки в любой момент времени $t\in[0,T]$

$$u(x,y,t)=\psi(x,y,t) \quad \text{при} \quad (x;y)\in\boldsymbol{\Gamma}_\Omega. \qquad (20.49)$$

Чтобы применить метод конечных разностей к поставленной смешанной задаче для уравнения теплопроводности, заданную пространственно-временную область $\Omega\times[0,T]$ покрываем ***сеткой*** $\Omega^{\tau}_{h^2}$, состоящей из ***узлов*** $(x_i;y_j;t_k)$ — точек параллелепипеда*), лежащих на пересечении трех плоскостей:

$$x=x_i, \quad \text{где} \quad x_i=ih, \quad i=0,1,\ldots,n, \quad h=\frac{1}{n},$$

$$y=y_j, \quad \text{где} \quad y_j=jh, \quad j=0,1,\ldots,n, \quad h=\frac{1}{n},$$

$$t=t_k, \quad \text{где} \quad t_k=k\tau, \quad k=0,1,\ldots,m, \quad \tau=\frac{T}{m}.$$

Узлы, соответствующие значениям индексов i, j, равным 0 или n, и значению $k=0$, считаются ***граничными***; остальные узлы — ***внутренние***. Узлы $(x_i;y_j;t_k)$, отвечающие одному фиксированному значению k, т.е. лежащие в фиксированной плоскости $t=t_k$, называются ***слоем***.

Аналогично обозначениям § 20.1 вводим следующие обозначения получающихся здесь сеточных функций:

$$u_{ij}^k :\approx u(x_i,y_j,t_k), \qquad f_{ij}^k := f(x_i,y_j,t_k),$$

$$\varphi_{ij} := \varphi(x_i,y_j), \qquad \psi_{ij}^k := \psi(x_i,y_j,t_k).$$

В соответствии с ними произведем сначала дискретизацию начального (20.48) и граничного (20.49) условий рассматриваемой задачи. Имеем:

во всех узлах нулевого слоя

$$u_{ij}^0=\varphi(x_i;y_j)=\varphi_{ij} \quad (i,j=0,1,\ldots,n); \qquad (20.50)$$

*) В целях упрощения записей ***шаги сетки*** по пространственным переменным x и y взяты одинаковыми: $h_1=h_2=h$. Для разных шагов см. упр. 20.7.

в граничных узлах произвольного k-го слоя

$$u_{0j}^k = \psi(0, y_j, t_k) = \psi_{0j}^k, \quad u_{nj}^k = \psi(1, y_j, t_k) = \psi_{nj}^k,$$
$$(j = 0, 1, \ldots, n); \qquad (20.51)$$

$$u_{i0}^k = \psi(x_i, 0, t_k) = \psi_{i0}^k, \quad u_{in}^k = \psi(x_i, 1, t_k) = \psi_{in}^k,$$
$$(i = 0, 1, \ldots, n). \qquad (20.52)$$

Дискретизацию уравнения (20.47) на сетке $\Omega_{h^2}^{\tau}$ будем производить с помощью шеститочечного шаблона, изображенного на рис. 20.8.

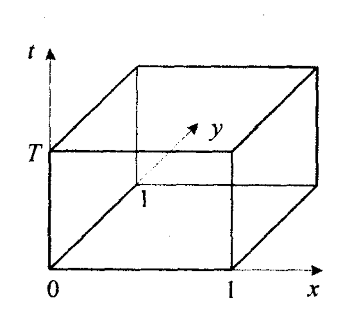


Рис. 20.7. Область задания уравнения (20.47)

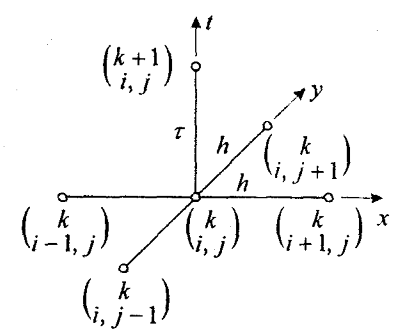


Рис. 20.8. Явный двухслойный шаблон для параболического уравнения с двумя пространственными переменными

Принимая точку $(x_i; y_j; t_k)$ за расчетный узел (внутренний по отношению к сторонам пластинки Ω при $i, j \neq 0, \neq n$), аппроксимируем в ней сначала оператор Лапласа

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2},$$

входящий в уравнение (20.47), на шаблоне типа «крест», служащем составной частью («основанием») рассматриваемого пространственного шаблона. Применив здесь формулу симметричной аппроксимации второй производной для той и другой пространственных переменных, получаем

$$\Delta u\Big|_{\substack{x=x_i\\ y=y_i\\ t=t_k}} = \frac{u_{i-1,\,j}^k - 2u_{ij}^k + u_{i+1,\,j}^k}{h^2} + \frac{u_{i,\,j-1}^k - 2u_{ij}^k + u_{i,\,j+1}^k}{h^2} + O(h^2), \quad (20.53)$$

т.е. в расчетном узле с точностью $O\left(h^2\right)$ значение выражения $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$ приближенно можно заменить значением

$$\Delta_h u_{ij}^k := \frac{1}{h^2}\left(u_{i-1,\,j}^k + u_{i+1,\,j}^k + u_{i,\,j-1}^k + u_{i,\,j+1}^k - 4u_{ij}^k\right). \qquad (20.54)$$

Далее, через значения функции u в узлах k-го и $(k+1)$-го слоев в соответствии с шаблоном можно записать

$$\frac{\partial u}{\partial t}\Big|_{\substack{x=x_i\\ y=y_i\\ t=t_k}} = \frac{u_{ij}^{k+1} - u_{ij}^k}{\tau} + O(\tau), \qquad (20.55)$$

т.е. с погрешностью $O(\tau)$ заменить в данном уравнении значение $\frac{\partial u}{\partial t}$ в расчетном узле значением

$$\frac{1}{\tau}(u_{ij}^{k+1} - u_{ij}^k). \qquad (20.56)$$

Таким образом, используя аппроксимации (20.53) и (20.55), т.е. подставляя в уравнение выражения (20.54) и (20.56), приходим к равенству

$$u_{ij}^{k+1} := u_{ij}^k + a^2 \frac{\tau}{h^2}\left(u_{i-1,\,j}^k + u_{i+1,\,j}^k + u_{i,\,j-1}^k + u_{i,\,j+1}^k - 4u_{ij}^k\right) + \tau f_{ij}^k, \quad (20.57)$$

представляющему собой явную формулу для вычисления приближенных значений неизвестной функции $u(x, y, t)$ в узловых точках $(k+1)$-го слоя по известным ее значениям на k-м слое. При этом, естественно, чтобы процесс вычислений мог быть начат и продолжен, привлекаются дополнительные условия (20.50)–(20.52) в граничных узлах.

Очевидно, что построенная ***явная разностная схема*** при достаточной гладкости решения аппроксимирует данную задачу с точностью $O\left(h^2 + \tau\right)$. Ее устойчивость доказана при условии

$$\tau \le \frac{h^2}{4a^2},$$

служащем существенным ограничением на шаг по времени, что приводит к большим вычислительным затратам при расчетах на сколько-нибудь протяженных временных промежутках T.

Для построения безусловно устойчивой схемы — аналога неявной схемы (20.19) — используем шаблон, симметричный изображенному на рис. 20.8 (вместо узла, помеченного как

$\binom{k+1}{i,\,j}$, $(k+1)$-го слоя берется симметричный ему относительно плоскости k-го слоя узел $\binom{k-1}{i,\,j}$ $(k-1)$-го слоя). При той же аппроксимации оператора Лапласа Δu в расчетной точке $(x_i;\,y_j;\,t_k)$ выражением $\Delta_h u_{ij}^k$ по формуле (20.54) и подобной (20.56) аппроксимации производной $\frac{\partial u}{\partial t}$ выражением

$$\frac{1}{\tau}\left(u_{ij}^k - u_{ij}^{k-1}\right)$$

с такой же, как и в (20.57), погрешностью аппроксимации $O\left(h^2+\tau\right)$ получаем ***неявную разностную схему***

$$u_{ij}^k - u_{ij}^{k-1} = a^2\frac{\tau}{h^2}\left(u_{i-1,\,j}^k + u_{i+1,\,j}^k + u_{i,\,j-1}^k + u_{i,\,j+1}^k - 4u_{ij}^k\right) + \tau f_{ij}^k,$$

которой с учетом обозначения

$$\gamma := a^2\frac{\tau}{h^2}, \tag{20.58}$$

придаем вид

$$\gamma u_{i-1,\,j}^k + \gamma u_{i+1,\,j}^k + \gamma u_{i,\,j-1}^k + \gamma u_{i,\,j+1}^k - (1+4\gamma)u_{ij}^k = -u_{ij}^{k-1} - \tau f_{ij}^k. \tag{20.59}$$

При каждом фиксированном $k = 1, 2, \ldots, m$ разностное уравнение (20.59) представляет собой СЛАУ относительно неизвестных значений сеточной функции u, принадлежащих k-му слою. Такая система при том или ином естественном упорядочивании неизвестных (сначала по i, потом по j, или наоборот) имеет пятидиагональную структуру (с «разнесенными» диагоналями) матрицы коэффициентов с очевидным диагональным преобладанием и в совокупности с дополнительными условиями (20.50)–(20.52) однозначно разрешима при любых значениях γ. Однако, прямое ее решение нельзя выполнить столь эффективно, как это делается в случае неявной схемы (20.19) для параболического уравнения с одной пространственной переменной, т.е. несмотря на наличие специальных методов решения сеточных уравнений (см. например, [161], добавление еще одной пространственной переменной принципиально увеличивает объем вычислений.

Встает вопрос о построении экономичных схем. Так называют аппроксимирующие задачу разностные схемы, безусловно

устойчивые и требующие при переходе от одного временного слоя к другому арифметических действий в количестве, имеющем один порядок с количеством узлов на слое [14]. Очевидно, неявные схемы с весами (20.24), в том числе чисто неявная схема (20.19) и схема Кранка–Николсон (20.25)–(20.26), построенные ранее для задачи теплопроводности с одной пространственной переменной, свойством экономичности обладают, чего нельзя сказать о схемах (20.57) и (20.59).

Для построения экономичной схемы решения задачи (20.47)–(20.49) используем следующий подход.

Между k-м и $(k+1)$-м слоями введем промежуточный слой расчетных точек $(x_i;\,y_j;\,t_{k+\frac{1}{2}})$, где $t_{k+\frac{1}{2}} := t_k + \frac{\tau}{2}$. Для подсчета значений

$$u_{ij}^{k+\frac{1}{2}} :\approx u\left(x_i;\,y_j;\,t_{k+\frac{1}{2}}\right)$$

во внутренних узлах промежуточного слоя в данном уравнении (20.47) аппроксимируем производные по формулам

$$\frac{\partial u}{\partial t} \approx \frac{u_{ij}^{k+\frac{1}{2}} - u_{ij}^k}{\tau/2},$$

$$\frac{\partial^2 u}{\partial x^2} \approx \frac{u_{i-1,\,j}^{k+\frac{1}{2}} - 2u_{ij}^{k+\frac{1}{2}} + u_{i+1,\,j}^{k+\frac{1}{2}}}{h^2},$$

$$\frac{\partial^2 u}{\partial y^2} \approx \frac{u_{i,\,j-1}^k - 2u_{ij}^k + u_{i,\,j+1}^k}{h^2}.$$

В результате получаем сеточное уравнение

$$u_{ij}^{k+\frac{1}{2}} = u_{ij}^k + \frac{a^2}{2}\frac{\tau}{h^2}\left(u_{i-1,\,j}^{k+\frac{1}{2}} - 2u_{ij}^{k+\frac{1}{2}} + u_{i+1,\,j}^{k+\frac{1}{2}} + u_{i,\,j-1}^k - 2u_{ij}^k + u_{i,\,j+1}^k\right) + \frac{\tau}{2}f_{ij}^{k+\frac{1}{2}},$$

где $f_{ij}^{k+\frac{1}{2}} := f(x_i,\,y_j,\,t_{k+\frac{1}{2}})$. Привлекая обозначение (20.58), это-

му уравнению придаем стандартный вид трехточечного уравнения второго порядка

$$u_{i-1,\,j}^{k+\frac{1}{2}} - \left(2+\frac{2}{\gamma}\right)u_{ij}^{k+\frac{1}{2}} + u_{i+1,\,j}^{k+\frac{1}{2}} = -\frac{\tau}{\gamma} f_{ij}^{k+\frac{1}{2}} - u_{i,\,j-1}^{k} + \left(2-\frac{2}{\gamma}\right)u_{ij}^{k} - u_{i,\,j+1}^{k}, \qquad (20.60)$$

связывающего неизвестные значения в трех соседних по направлению оси Ox узловых точках промежуточного слоя.

Так как правую часть равенства (20.60) при переходе от k-го к $(k+1)$-му слою можно считать известной, то при каждом фиксированном $j = 1, 2, \ldots, n-1$ оно представляет собой систему из $n-1$ уравнений с $n+1$ неизвестными, два из которых фактически известны из граничных условий (20.51). Матрицы коэффициентов всех таких $n-1$ СЛАУ, соответствующих различным значениям j, одинаковые трехдиагональные с явным диагональным преобладанием при любом значении γ. Соответственно, каждая из них может быть решена методом прогонки за $O(n)$ число арифметических операций (см. § 2.6), т.е. для нахождения $(n-1)^2$ значений неизвестных на промежуточном слое количество требуемых арифметических операций составит величину $O(n^2)$.

Переход от $\left(k+\frac{1}{2}\right)$-го слоя к $(k+1)$-му совершается на основе аппроксимации производных по формулам

$$\frac{\partial u}{\partial t} \approx \frac{u_{ij}^{k+1} - u_{ij}^{k+\frac{1}{2}}}{\tau/2},$$

$$\frac{\partial^2 u}{\partial x^2} \approx \frac{u_{i-1,\,j}^{k+\frac{1}{2}} - 2u_{ij}^{k+\frac{1}{2}} + u_{i+1,\,j}^{k+\frac{1}{2}}}{h^2},$$

$$\frac{\partial^2 u}{\partial y^2} \approx \frac{u_{i,\,j-1}^{k+1} - 2u_{ij}^{k+1} + u_{i,\,j+1}^{k+1}}{h^2}.$$

подстановка которых в исходное уравнение (условно для той же расчетной точки $(x_i; y_j; t_{k+\frac{1}{2}})$) приводит к аналогичному (20.60)

уравнению

$$u_{i,\,j-1}^{k+1} - \left(2+\frac{2}{\gamma}\right)u_{ij}^{k+1} + u_{i,\,j+1}^{k+1} = -\frac{\tau}{\gamma} f_{ij}^{k+\frac{1}{2}} - u_{i-1,\,j}^{k+\frac{1}{2}} + \left(2-\frac{2}{\gamma}\right)u_{ij}^{k+\frac{1}{2}} - u_{i+1,\,j}^{k+\frac{1}{2}}, \qquad (20.61)$$

Фиксируя здесь $i = 1, 2, \ldots, n-1$, будем получать линейные системы относительно неизвестных узловых значений вдоль направления оси Oy, причем матрицы этих систем — те же, что и в предыдущем случае, следовательно, привлекая граничные значения (20.52), методом прогонки все искомые значения на $(k+1)$-м слое могут быть подсчитаны за $O(n^2)$ арифметических операций.

Итак, совокупность формул (20.60), (20.61) и дополнительных к ним начальных и граничных условий (20.50)–(20.52) определяет экономичную неявную разностную схему, которая, как доказано [78, 92], является абсолютно устойчивой и аппроксимирует исходную задачу (20.47)–(20.49) со вторым порядком по каждой из переменных x, y, t. Метод решения параболических задач с двумя пространственными переменными, основанный на построении подобных схем (называемых ***продольно-поперечными схемами***), носит названия ***метод переменных направлений***, ***метод продольно-поперечных прогонок***, ***метод Писмэна-Рэчфорда****) [204]. Это один из лучших методов решения таких задач [78], однако, он не имеет непосредственного обобщения на случай трех пространственных переменных в параболическом уравнении.

Более продуктивными в плане обобщений являются ***методы расщепления***. В наиболее абстрактной постановке речь идет о расщеплении операторов [13, 14, 118], в менее абстрактной — можно говорить о расщеплении по физическим процессам и по координатам.

Рассмотрим вкратце один из вариантов ***метода покоординатного расщепления***, который также можно назвать ***методом дробных шагов*** (впрочем, как и метод переменных направлений).

Представив функцию f в виде $f_1 + f_2$, данное уравнение (20.47) заменяем двумя:

$$\frac{1}{2}\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2} + f_1(x, y, t)$$

и

---

*) В разных переводах можно встретить написание *Рэчфорд*, *Речфорд*, *Рекфорд*, *Рэкфорд*.

$$\frac{1}{2}\frac{\partial u}{\partial t} = a^2\frac{\partial^2 u}{\partial y^2} + f_2(x, y, t),$$

в сумме дающими (20.47). Эти два уравнения аппроксимируем соответственно сеточными уравнениями

$$\frac{u_{ij}^{k+\frac{1}{2}} - u_{ij}^{k}}{\tau} = a^2\frac{u_{i-1,j}^{k+\frac{1}{2}} - 2u_{ij}^{k+\frac{1}{2}} + u_{i+1,j}^{k+\frac{1}{2}}}{h^2} + (f_1)_{ij}^{k+\frac{1}{2}}, \qquad (20.62)$$

$$\frac{u_{ij}^{k+1} - u_{ij}^{k+\frac{1}{2}}}{\tau} = a^2\frac{u_{i,j-1}^{k+1} - 2u_{ij}^{k+1} + u_{i,j+1}^{k+1}}{h^2} + (f_2)_{ij}^{k+1}, \qquad (20.63)$$

в которых узнаём чисто неявные схемы для одномерного уравнения теплопроводности. Решая прогонкой разностное уравнение (20.62) (СЛАУ с трехдиагональной матрицей коэффициентов при каждом фиксированном j), находим значения $u_{ij}^{k+\frac{1}{2}}$ на промежуточном слое, затем из (20.63) последовательным фиксированием i аналогично находим все значения u_{ij}^{k+1} неизвестных на $(k+1)$-м слое, что требует всего $O(n^2)$ арифметических операций и означает экономичность совокупной разностной схемы (20.62)–(20.63).

Особенностью такого ***локально-одномерного метода*** является то обстоятельство, что ни одно из разностных уравнений (20.62), (20.63) по отдельности не аппроксимирует исходное уравнение (20.47), но в совокупности они обеспечивают эту аппроксимацию с точностью $O(h^2 + \tau)$, в связи с чем такой метод относят к ***методам суммарной аппроксимации***. Метод абсолютно устойчив.

УПРАЖНЕНИЯ

20.1. А) Получите таблицу, реализующую ε-схему (см. § 20.2) разностного уравнения (20.20) при $\gamma = 0.5$, $g \equiv 0$ на трех слоях. Убедитесь в том, что значение u_i^{k+3} содержит ошибку -24ε при одиночной ошибке значения u_i^k величиной ε [20].

Б) Проведите на нескольких слоях ε-анализ разностной схемы (20.20) с $g_i \equiv 0$ и произвольными $\gamma > 0$. Существуют ли значения γ, при которых можно рассчитывать на устойчивость схемы?

20.2. Проведя ε-анализ явной двухслойной схемы (20.18) при $g_i \equiv 0$, ответьте, противоречит ли его результат условию устойчивости (20.22)?

20.3. Выведите формулу (20.40), привлекая разложение $u(x_i, t)$ по формуле Тейлора в точке $(x_i; 0)$ и заменяя $\dfrac{\partial^2 u(x_i, 0)}{\partial t^2}$ выражением $a^2\dfrac{\partial^2 u(x_i, 0)}{\partial x^2}$.

20.4. Запишите явную трехслойную разностную схему, аппроксимирующую с точностью $O(h^2 + \tau^2)$ уравнение в ы н у ж д е н н ы х колебаний однородной струны, начальные и краевые условия для которого задаются формулами (20.30)–(20.32).

20.5. Пользуясь шаблоном, изображенным на рис. 20.9, постройте неявную схему, аппроксимирующую гиперболическую задачу (20.29)–(20.32) с точностью $O(h^2 + \tau^2)$ [178]. Запишите алгоритм, реализующий получение каркаса приближенного решения с помощью этой схемы.

Рис. 20.9

20.6. Составьте явную разностную схему, соответствующую шаблону рис. 20.10а и аппроксимирующую задачу

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = f(x, t) \quad \text{при} \quad (x; t) \in [0, 1]\times[0, T],$$

$$u(x, 0) = \varphi(x) \quad \text{при} \quad x \in [0, 1],$$

$$u(0, t) = \psi(t) \quad \text{при} \quad t \in [0, T]$$

(где $c > 0$, $\varphi(0) = \psi(0)$) с первым порядком точности по x и по t.

Можно ли считать пригодной аналогичную схему, опирающуюся на шаблон рис. 20.10б ? Почему?

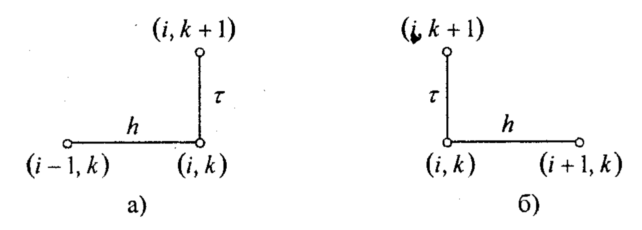


Рис. 20.10

20.7. А) Запишите расчетные формулы явной двухслойной разностной схемы типа схемы (20.57) для случая, когда смешанная задача теплопроводности (20.47)–(20.49) ставится для прямоугольной пластинки $\Omega := [0, l_1] \times [0, l_2]$ и шаги сетки по x и по y неодинаковые (h_1 и h_2, соответственно).

Б) Для указанного в А случая запишите формулы продольно-поперечной схемы.

20.8. Изобразите шаблон, соответствующий неявной двухслойной разностной схеме (20.59) для задачи (20.47)–(20.49). Выполните линейное упорядочивание узлов первого слоя при шаге $h = 0.25$ и покажите структуру соответствующей этому слою матрицы СЛАУ:

а) в случае однородного граничного условия (20.49) ($\psi \equiv 0$);

б) в случае неоднородного граничного условия ($\psi \not\equiv 0$).

20.9. Обобщите задачу теплопроводности (20.47)–(20.49) на случай трех пространственных переменных.

А) Составьте для нее явную разностную схему (основой для которой должен служить аналог формулы (20.57)).

Б) Запишите для нее расчетные формулы локально-одномерного метода (типа формул (20.62)–(20.63)).

ГЛАВА 21 МЕТОД КОНЕЧНЫХ РАЗНОСТЕЙ ДЛЯ СТАЦИОНАРНЫХ ЗАДАЧ

В области задания двумерного эллиптического уравнения вводится равномерная прямоугольная сетка, строятся разностные схемы второго порядка точности (на шаблоне типа «крест»), показываются приемы учета граничных условий первого рода в случае произвольной области и способы аппроксимации нормальных производных в граничных условиях второго и третьего рода. Подчеркивается разреженность матриц получающихся при этом систем сеточных уравнений и, в частности, их блочно-трехдиагональная структура в случае, когда исходная задача ставится в прямоугольнике. Обсуждаются особенности и целесообразность применения к таким специфичным системам изучавшихся в главах 2, 3 прямых и итерационных методов решения СЛАУ (Гаусса, МПИ, Зейделя, ПВР) и указывается на наличие хорошо приспособленных для этого экономичных методов, среди которых выделяется итерационный метод переменных направлений, реализующий принцип установления. Рассматривается численный пример (с использованием методов Зейделя и ПВР).

21.1. КОНЕЧНОРАЗНОСТНАЯ ДИСКРЕТИЗАЦИЯ КРАЕВЫХ ЗАДАЧ ДЛЯ ЭЛЛИПТИЧЕСКИХ УРАВНЕНИЙ

Специфичность применения метода конечных разностей к стационарным задачам проявляется, в основном, на двух этапах: при аппроксимации граничных условий и при решении получающихся в итоге дискретных задач большой размерности. В этом параграфе будем рассматривать лишь формальное построение разностных схем для различных постановок стационарных задач на основе замены производных «стандартными» разностными отношениями и учета граничных условий тем или иным способом.

Обратимся сначала к наиболее простой стационарной задаче. А именно, построим разностную схему, отвечающую задаче Дирихле для уравнения Пуассона в области Ω с границей $\mathbf{\Gamma}$ в случае, когда Ω — прямоугольник $[a, b] \times [c, d]$:

$$\Delta u := \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y) \quad \text{при} \quad (x; y) \in \Omega, \tag{21.1}$$

$$u(x, y) = \varphi(x, y) \quad \text{при} \quad (x; y) \in \mathbf{\Gamma} \tag{21.2}$$

(см. (19.5), (19.14)).

Покроем данную двумерную область Ω ***сеткой узлов*** $(x_i; y_j)$, образованных пересечениями прямых $x = x_i$ и $y = y_j$, где

$$x_i = a + ih_1, \quad h_1 = \frac{b-a}{n}, \quad i = 0, 1, \dots, n,$$
$$y_j = c + jh_2, \quad h_2 = \frac{d-c}{m}, \quad j = 0, 1, \dots, m \qquad (21.3)$$

(h_1 и h_2 — ***шаги сетки*** по осям Ox и Oy соответственно). Будем называть узлы $(x_i; y_j)$ ***внутренними***, когда $i \in \{1, 2, \dots, n-1\}$, $j \in \{1, 2, \dots, m-1\}$, и ***граничными***, когда $i = 0$ или $i = n$, а $j \in \{1, 2, \dots, m-1\}$, и когда $j = 0$ или $j = m$, а $i \in \{1, 2, \dots, n-1\}$ (см. рис. 21.1, где внутренние узлы помечены кружочками, а граничные — крестиками). Множество внутренних узлов сетки обозначаем Ω_{h_1, h_2}, граничных узлов — $\mathbf{\Gamma}_{h_1, h_2}$, всех узлов (внутренних и граничных) — $\overline{\Omega}_{h_1, h_2}$. Заметим, что узлы $(x_0; y_0)$, $(x_n; y_0)$, $(x_0; y_m)$, $(x_n; y_m)$, соответствующие вершинам данного прямоугольника Ω, при использовании шаблона типа «крест» (помечен на рисунке темными кружками) в расчетах не участвуют и не относятся ни к внутренним, ни к граничным узлам.

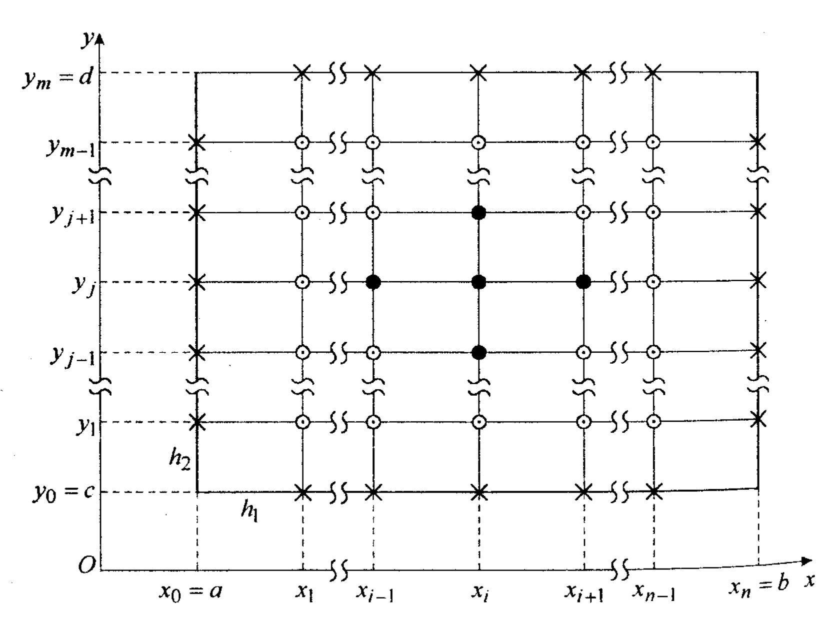


Рис. 21.1. *Сетка и шаблон для дискретизации задачи* (21.1)–(21.2)

Рассматривая данное уравнение (21.1) в расчетной точке $(x_i; y_j) \in \Omega_{h_1, h_2}$ и привлекая в соответствии с указанным пятиточечным шаблоном «крест» формулы симметричной аппроксимации вторых производных для получения дискретной версии оператора Лапласа Δu (по аналогии с формулой (20.53) в § 20.6), в предположении о достаточной гладкости решения $u(x, y)$ (четырехкратной непрерывной дифференцируемости по x и по y) имеем

$$\frac{u(x_{i-1}, y_j) - 2u(x_i, y_j) + u(x_{i+1}, y_j)}{h_1^2} + O(h_1^2) +$$
$$\frac{u(x_i, y_{j-1}) - 2u(x_i, y_j) + u(x_i, y_{j+1})}{h_2^2} + O(h_2^2) = f(x_i, y_j).$$

Отбрасывая здесь остаточные члены $O(h_1^2)$, $O(h_2^2)$ и используя обозначения

$$f_{ij} := f(x_i, y_j), \quad u_{ij} :\approx u(x_i, y_j), \qquad (21.4)$$

приходим к разностному уравнению

$$\frac{1}{h_1^2}(u_{i-1, j} - 2u_{ij} + u_{i+1, j}) + \frac{1}{h_2^2}(u_{i, j-1} - 2u_{ij} + u_{i, j+1}) = f_{ij} \qquad (21.5)$$

(где $i = 1, 2, \dots, n-1$; $j = 1, 2, \dots, m-1$), которыми подменяем данное уравнение на сетке Ω_{h_1, h_2}.

Формально в полученном таким образом разностном уравнении (21.5) неизвестных столько, сколько всего узлов содержится в множестве $\overline{\Omega}_{h_1, h_2}$, т.е. $(n+1)\times(m+1) - 4$. Однако часть из них (конкретно, $2m + 2n - 4$) сразу определяется точно из граничного условия (21.2):

$$\left.\begin{aligned} u_{0j} &= u(x_0, y_j) = \varphi(x_0, y_j) =: \varphi_{0j}, \\ u_{nj} &= u(x_n, y_j) = \varphi(x_n, y_j) =: \varphi_{nj} \end{aligned}\right| \quad j = 1, 2, \dots, m-1;$$
$$\left.\begin{aligned} u_{i0} &= u(x_i, y_0) = \varphi(x_i, y_0) =: \varphi_{i0}, \\ u_{in} &= u(x_i, y_n) = \varphi(x_i, y_n) =: \varphi_{in} \end{aligned}\right| \quad i = 1, 2, \dots, n-1. \qquad (21.6)$$

Следовательно, дополнив (21.5) совокупностью сеточных граничных условий (21.6), получаем СЛАУ с квадратной матрицей коэффициентов относительно значений каркаса приближенного решения данной задачи Дирихле во внутренних узлах сетки.

Построенная разностная схема (21.5)–(21.6) аппроксимирует задачу (21.1)–(21.2) со вторым порядком точности и устойчива.

Последнее, как правило, доказывается с помощью разностного (сеточного) принципа максимума [14, 92, 100, 103, 153, 154, 155, 158, 159].

Не вызывает затруднений формальная аппроксимация разностной схемой со вторым порядком точности на той же сетке $\overline{\Omega}_{h_1,h_2}$ с тем же шаблоном «крест» более общего уравнения

$$A\frac{\partial^2 u}{\partial x^2}+C\frac{\partial^2 u}{\partial y^2}+D\frac{\partial u}{\partial x}+E\frac{\partial u}{\partial y}+Gu=F \qquad (21.7)$$

(сравните с (19.13)), если оно задано в том же прямоугольнике $\Omega=[a,b]\times[c,d]$ с тем же граничным условием (21.2), когда коэффициенты A, C, D, E, G и свободный член F уравнения зависят от переменных x и y, причем A и C в Ω удовлетворяют условию эллиптичности $AC>0$.

Действительно, используя для вторых производных те же аппроксимации, что и выше, аппроксимируем первые производные на том же шаблоне по симметричным формулам

$$\left.\frac{\partial u}{\partial x}\right|_{\substack{x=x_i\\ y=y_j}}=\frac{u(x_{i+1},y_j)-u(x_{i-1},y_j)}{2h_1}+O(h_1^2),$$

$$\left.\frac{\partial u}{\partial y}\right|_{\substack{x=x_i\\ y=y_j}}=\frac{u(x_i,y_{j+1})-u(x_i,y_{j-1})}{2h_2}+O(h_2^2).$$

Тогда, фиксируя в уравнении (21.7) точку $(x;y)=(x_i;y_j)$ и привлекая обозначения (21.4) и им аналогичные, после отбрасывания остаточных членов получаем соответствующее (21.7) сеточное уравнение

$$A_{ij}\frac{u_{i-1,j}-2u_{ij}+u_{i+1,j}}{h_1^2}+C_{ij}\frac{u_{i,j-1}-2u_{ij}+u_{i,j+1}}{h_2^2}+D_{ij}\frac{u_{i+1,j}-u_{i-1,j}}{2h_1}+$$

$$+E_{ij}\frac{u_{i,j+1}-u_{i,j-1}}{2h_2}+G_{ij}u_{ij}=F_{ij}. \qquad (21.8)$$

Легко видеть, что уравнение (21.8) не имеет принципиальных отличий от уравнения (21.5) в том плане, что также представляет собой пятиточечное разностное уравнение (связывающее пять значений в узлах «креста» — значения $u_{i-1,j}$, $u_{i+1,j}$, u_{ij}, $u_{i,j-1}$, $u_{i,j+1}$) и при всевозможных значениях $i\in\{1,2,\ldots,n-1\}$, $j\in\{1,2,\ldots,m-1\}$ — это СЛАУ с числом уравнений, совпадающим с числом неизвестных, если дополнить ее сеточными граничными условиями (21.6), и с аналогичной структурой матрицы коэффициентов. Поэтому далее ограничимся рассмотрением

простейшего эллиптического уравнения (21.1), варьируя лишь постановки краевых задач для него.

Построение разностной схемы для эллиптического уравнения несколько усложняется, когда первая краевая задача для него ставится в области Ω, отличной от прямоугольной.

Пусть Ω в задаче (21.1)–(21.2) — произвольная конечная односвязная область с непрерывной границей Γ, определяемой некоторым уравнением $\psi(x,y)=0$. Поместим эту область в прямоугольник $[x_{\min},x_{\max}]\times[y_{\min},y_{\max}]$ (рис. 21.2), и на этот прямоугольник наложим сетку (21.3), где принимаем

$$a:=x_{\min},\quad b:=x_{\max},\quad c:=y_{\min},\quad d:=y_{\max}.$$

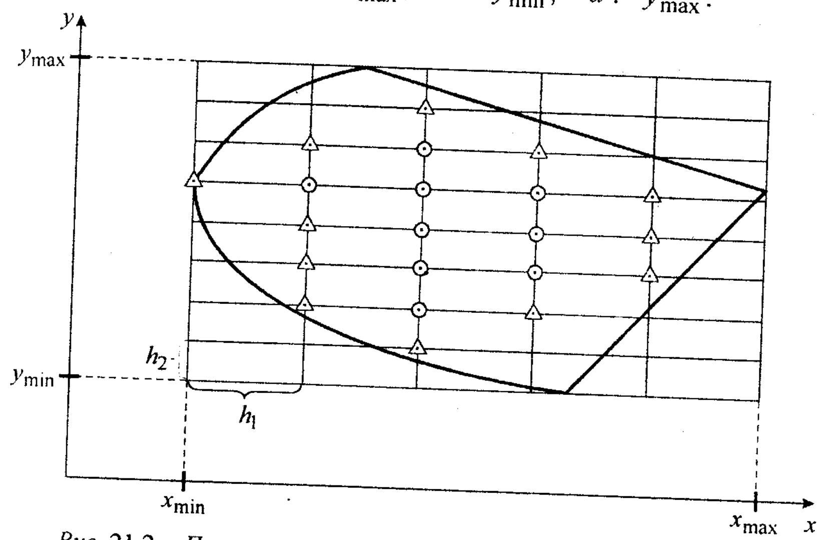


Рис. 21.2. Прямоугольная сетка для непрямоугольной области

Разбиение множества всех узлов, попадающих в область Ω или на ее границу, на внутренние и граничные здесь уже не столь однозначно, как это было в случае прямоугольной области. Имеется несколько различных подходов к этому, и в каждом из них такое разбиение обычно связывают с выбранным для аппроксимации уравнения шаблоном. В случае «креста» наиболее простой подход — считать узлы $(x_i;y_j)$ внутренними, если они принадлежат области $\overline{\Omega}:=\Omega\cup\Gamma$ вместе со своими четырьмя «соседями» — вершинами «креста» с центром в точке $(x_i;y_j)$, и граничными, если хотя бы один из этих «соседей» лежит за пределами $\overline{\Omega}$ [92].

Примем за основу следующую, более естественную, на наш взгляд, классификацию узлов.

Узлы определяемой равенствами (21.3) сетки будем называть ***внутренними***, если они принадлежат данной открытой области Ω (множество их обозначаем Ω_{h_1,h_2}), и ***граничными***, если они попадают на границу $\mathbf{\Gamma}$ области Ω (множество таких узлов $\mathbf{\Gamma}_{h_1,h_2}$). Узлы $(x_i; y_j) \in \Omega_{h_1,h_2}$, служащие центрами «крестов», целиком (т. е. вместе со всеми четырьмя вершинами) входящих во множество $\Omega_{h_1,h_2} \cup \mathbf{\Gamma}_{h_1,h_2}$, называем ***строго внутренними***; всю их совокупность обозначаем $\Omega^0_{h_1,h_2}$. Узлы, служащие вершинами «крестов» с центрами в строго внутренних узлах и не являющиеся при этом строго внутренними, называем ***приграничными узлами***; их множество $\mathbf{\Gamma}^0_{h_1,h_2}$ образует так называемую ***граничную полоску*** (см. рис. 21.2, где строго внутренние узлы обозначены кружками, а приграничные — треугольниками). Очевидно, что могут быть такие узлы, которые принадлежат $\overline{\Omega}$ и при этом не относятся ни к строго внутренним, ни к приграничным (на рис. 21.2 таких узлов три, найдите их!).

Для задачи (21.1)–(21.2) в произвольной области Ω каждому строго внутреннему узлу $(x_i; y_j)$ ставится в соответствие одно сеточное уравнение вида (21.5). Множество таких уравнений нужно дополнить сеточными уравнениями относительно неизвестных значений приближенного решения в приграничных узлах, используя для этого каким-нибудь образом граничное условие (21.2). Наиболее простой способ сделать это — осуществить ***простой снос***, т.е. заменить значение $u(x, y)$ в каждой точке граничной полоски значением данной функции $\varphi(x, y)$ в ближайшей (по x, по y или по нормали к границе $\mathbf{\Gamma}$) точке границы $\mathbf{\Gamma}$. Однако, следует иметь в виду, что такой грубый учет граничных условий может значительно повлиять на качество аппроксимации задачи в целом. Несколько увеличить точность аппроксимации, не отказываясь от простого сноса, можно за счет расширения множества используемых в расчетах узлов: в качестве приграничных узлов использовать «заграничные», находящиеся от границы на расстоянии, меньшем полушага $0.5h_1$ или $0.5h_2$, а соответствующие приграничные внутренние узлы перевести в разряд строго внутренних.

Альтернативой простому сносу при аппроксимации граничных условий может служить предложенный немецким математиком Л. Коллатцем ***способ линейной интерполяции***. Заключается он в следующем.

Пусть: P — приграничный узел, M — ближайший к нему строго внутренний узел, N — ближайшая к ним точка границы $\mathbf{\Gamma}$ области Ω; все три точки лежат на одной прямой $x = x_i$ или $y = y_j$; известны расстояния $|MP| = h$, $|PN| = q$ (рис. 21.3). Естественно предположить, что значение искомой функции u в точке P

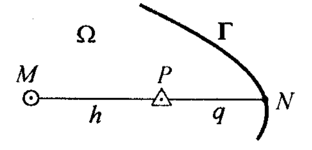


Рис. 21.3. К учету граничного условия способом линейной интерполяции

приближенно линейно зависит от значений $u(M)$ и $u(N) = \varphi(N)$, и его можно получить, разделив эти значения в отношении $\lambda := \frac{h}{q}$, в котором точка P делит отрезок MN. Тогда по известной формуле деления отрезка в данном отношении можно принять

$$u(P) \approx \frac{u(M) + \lambda u(N)}{1 + \lambda} = \frac{qu(M) + h\varphi(N)}{q + h}.$$

Если простой снос вносит погрешность аппроксимации $O(h)$, то для способа линейной интерполяции погрешность аппроксимации граничных условий составляет величину $O(h^2)$ [100].

Имеются и другие способы аппроксимации граничных условий в задачах Дирихле (см. [14] и др.).

При решении методом конечных разностей в т о р о й краевой задачи для эллиптического уравнения, оставляя без изменений аппроксимацию самого дифференциального уравнения разностным на той же прямоугольной сетке $\Omega^0_{h_1,h_2}$ строго внутренних узлов, при получении дополнительных связей в приграничных узлах нужно аппроксимировать в них граничное условие вида (см. (19.15))

$$\left.\frac{\partial u}{\partial \mathbf{n}}\right|_{(x;\, y) \in \mathbf{\Gamma}} = \varphi(x, y), \tag{21.9}$$

где $\mathbf{n}$ — направление внешней нормали к границе $\mathbf{\Gamma}$ области Ω. Сделать это можно, например, следующим образом.

Рассмотрим получение дополнительного к основным разностного уравнения на основе заданного равенства (21.9) в приграничном узле P в ситуации, изображенной на рис. 21.4.

Выполнив ортогональное (или какое-либо другое) проектирование точки P на границу Γ, приходим к точке Q, в которой, в силу условия (21.9), имеем

$$\frac{\partial u(Q)}{\partial \mathbf{n}} = \varphi(Q),$$

а исходя из близости точек P и Q, полагаем

$$\frac{\partial u(P)}{\partial \mathbf{n}} \approx \varphi(Q)$$

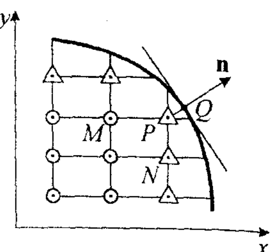


Рис. 21.4. К получению разностного уравнения в приграничном узле P

Считая, что нормаль $\mathbf{n}$ в рассматриваемой точке составляет угол α с осью Ox, можно записать выражение производной по направлению нормали $\mathbf{n}(\cos\alpha; \sin\alpha)$ через частные производные по координатам:

$$\frac{\partial u(P)}{\partial \mathbf{n}} = \frac{\partial u(P)}{\partial x}\cos\alpha + \frac{\partial u(P)}{\partial y}\sin\alpha .$$

Частные производные, в свою очередь, аппроксимируем простейшими разностными отношениями через значения неизвестных в соседних узлах M и N. Таким образом, для приграничного узла $P(x_l; y_k)$ имеем приближенное равенство

$$\frac{u(x_l, y_k) - u(x_{l-1}, y_k)}{h_1}\cos\alpha + \frac{u(x_l, y_k) - u(x_l, y_{k-1})}{h_2}\sin\alpha \approx \varphi(Q),$$

которое с учетом $u_{ij} \approx u(x_i; y_j)$ превращаем в разностное уравнение

$$\frac{\cos\alpha}{h_1}(u_{lk} - u_{l-1,\,k}) + \frac{\sin\alpha}{h_2}(u_{lk} - u_{l,\,k-1}) = \varphi(Q). \qquad (21.10)$$

Для получения разностной схемы, аппроксимирующей, например, задачу Неймана (21.1), (21.9) с точностью $O(h)$, к основным сеточным уравнениям вида (21.5) должно быть добавлено столько уравнений вида (21.10), сколько узлов содержится в граничной полоске $\Gamma^0_{h_1, h_2}$.

Как видим, в случае произвольной области Ω конечноразностная аппроксимация второй краевой задачи является делом, гораздо более хлопотным, чем первой.

Если какая-то часть границы Γ параллельна координатной оси Ox или Oy, то в принадлежащих этой части границы точках производная по направлению нормали будет являться просто частной производной по соответствующей переменной x или y.

В таком случае, получение «граничных» разностных уравнений типа уравнения (21.10) и сам их вид упрощаются (особенно, если параллельная оси часть границы включается в число прямых, с помощью которых образуется сетка). Точность аппроксимации при этом можно на порядок повысить, если в такой ситуации применить ***метод фиктивной точки***, использовавшийся ранее (§ 20.4) при аппроксимации дифференциального начального условия в задаче о свободных колебаниях струны.

Пусть часть границы Γ («правая») проходит по прямой $x = b$, включенной в семейство прямых $x = x_i$ при $i = n$. Беря за расчетный узел граничную точку $(x_n; y_j)$, дополняем этот узел, граничные узлы $(x_n; y_{j-1})$, $(x_n; y_{j+1})$ и внутренний узел $(x_{n-1}; y_j)$ до «креста» фиктивным узлом $(x_{n+1}; y_j)$, взятым за пределами Ω симметрично внутреннему узлу $(x_{n-1}; y_j)$ (рис. 21.5). На этом шаблоне проводим аппроксимацию заданного дифференциального уравнения; для задачи Неймана (21.1), (21.9) такой аппроксимацией служит сеточное уравнение

$$\frac{1}{h_1^2}(u_{n-1,\,j} - 2u_{nj} + u_{n+1,\,j}) + \frac{1}{h_2^2}(u_{n,\,j-1} - 2u_{nj} + u_{n,\,j+1}) = f_{nj}. \quad (21.11)$$

Условие (21.9) для этого случая можно записать в виде

$$\left.\frac{\partial u}{\partial x}\right|_{\substack{x=x_n\\ y=y_j}} = \varphi(x_n, y_j)$$

и аппроксимировать с точностью $O(h_1^2)$ разностным уравнением

$$\frac{u_{n+1,\,j} - u_{n-1,\,j}}{2h_1} = \varphi_{nj}.$$

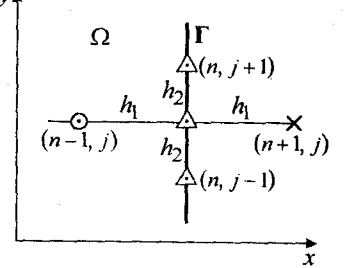


Рис. 21.5. К методу фиктивной точки («правый вертикальный» участок границы Γ)

Выражая отсюда фиктивное значение $u_{n+1,j}$, исключаем его из уравнения (21.11), в результате чего приходим к соответствующему взятой за основу граничной точке разностному уравнению второго порядка точности

$$\frac{2}{h_1^2}(u_{n-1,\,j} - u_{nj}) + \frac{1}{h_2^2}(u_{n,\,j-1} - 2u_{nj} + u_{n,\,j+1}) = f_{nj} - \frac{2}{h_1}\varphi_{nj}. \quad (21.12)$$

Построение разностных схем для третьей краевой задачи не привносит ничего нового по сравнению с методикой построения разностных схем для второй краевой задачи.

21.2. О СПЕЦИФИКЕ СЛАУ, АППРОКСИМИРУЮЩИХ ЭЛЛИПТИЧЕСКИЕ УРАВНЕНИЯ, И ПРЯМЫХ МЕТОДАХ ИХ РЕШЕНИЯ

Как следует из предыдущего параграфа, применение метода конечных разностей с пятиточечным шаблоном типа «крест» к краевым задачам для эллиптических уравнений сводит их к решению систем алгебраических уравнений вида

$$a_{ij}u_{i-1,\,j} + b_{ij}u_{i+1,\,j} + c_{ij}u_{i,\,j-1} + d_{ij}u_{i,\,j+1} + e_{ij}u_{ij} = f_{ij} \qquad (21.13)$$

(сравните с (21.5), (21.8), где i и j пробегают независимо множества целочисленных значений от 1 до $n-1$ и до $m-1$ соответственно). В зависимости от заданных граничных условий уравнения (21.13) дополняются теми или иными значениями неизвестных в граничных (или приграничных) узлах или вспомогательными уравнениями с участием этих неизвестных (см., например, (21.6), (21.12)), что обеспечивает соответствие числа уравнений в итоговой алгебраической системе числу фигурирующих в (21.13) неизвестных. При определенных требованиях к решаемой дифференциальной задаче, таких, как требование $u(x, y) \in C^4(\overline{\Omega})$, можно гарантировать, что заменяющая ее рассмотренным способом дискретная (сеточная) задача при должной аппроксимации граничных условий аппроксимирует данную задачу со вторым порядком точности*), однозначно разрешима, и имеет место сходимость (с тем же порядком) решений этой дискретной задачи к решению исходной задачи в процессе бесконечного сгущения сетки по обоим направлениям одновременно, т.е. по x и по y.

Наличие сходимости означает, что чем меньше расчетный шаг h (будем для простоты считать, что используется сетка с квадратными ячейками, т.е. $h_1 = h_2 =: h$), тем точнее сеточное решение $\{u_{ij}\}$ представляет точное решение $u(x, y)$. Таким образом, желание получить более точное сеточное приближение к истинному решению исходной задачи порождает необходимость в уменьшении шага h. Однако такое уменьшение шага вызывает существенное увеличение количества уравнений в системе (21.13) (оно обратно пропорционально квадрату шага), что, в свою очередь, приводит к большому росту вычислительных затрат на ее решение.

Поскольку использование оценок погрешностей аппроксимаций в реальных задачах почти невозможно, здесь часто применяют упоминавшийся ранее (§§12.5, 14.7) ***принцип Рунге***, сводящийся к тому, что проводя вычисления на сгущающихся сетках, сравнивают значения решения сеточной задачи в общих для этих сеток узлах.

Конкретнее, если, например, проведены расчеты на квадратной сетке с некоторым фиксированным шагом h и получено сеточное решение u_h, затем шаг уменьшен вдвое и на сетке с шагом $0.5h$ получено решение $u_{0.5h}$, и если выполняется неравенство

$$\|u_h - u_{0.5h}\| \le 3\varepsilon, \qquad (21.14)$$

где ε — заданная точность для сеточного решения в используемой сеточной норме, то в соответствии с правилом Рунге, каркас решения $u_{0.5h}$ считают верным с точностью ε благодаря наличию теоретической оценки погрешности аппроксимации $O(h^2)$.

Но принцип Рунге «работает», когда кроме всех прочих условий, оправдывающих его применение, сеточное уравнение решается с небольшей, чем ε, погрешностью. К сожалению, в процессе $h \to 0$ ухудшается обусловленность систем сеточных уравнений, что приводит к дисбалансу между желаемой точностью аппроксимации и, возможно, чрезмерно большой погрешностью решения таких систем. Это означает, что могут встретиться ситуации, когда при наличии формальных гарантий сходимости в процессе сгущения сетки неравенство (21.14) может вообще не выполниться, если ε окажется слишком малым, так как в реальном процессе дробления шага величины $\|u_h - u_{0.5h}\|$ сначала убывают, отражая факт сходимости, а далее начинают возрастать под влиянием ошибок округления при решении все хуже обусловленных СЛАУ все большей размерности.

Обсудим теперь вопрос о том, какие методы могут быть применены для решения систем сеточных уравнений вида (21.13). Будем, опять таки, для простоты считать, что сетка имеет квадратные ячейки с шагом h и что общее число определяемых неизвестных (составляющих каркас искомого приближенного решения после исключения его граничных значений) есть*)

$$N := (n-1)(m-1). \qquad (21.15)$$

Тогда если всю эту $(n-1)\times(m-1)$-матрицу неизвестных u_{ij} тем

*) Существуют схемы и более высокого порядка точности, опирающиеся на другие шаблоны, например, на девятиточечный [13, 14, 154].

*) Определенное в (21.15) число N используем в рассуждениях и с произвольными (неквадратными) сетками.

или иным способом упорядочить линейным образом, т.е. записать в виде N-мерного вектора $\mathbf{u}_h$, например, так*):

$$\mathbf{u}_h := (u_{11}; u_{21}; \ldots; u_{n-1,1}; \quad u_{12}; u_{22}; \ldots; u_{n-1,2}; \quad \ldots$$

$$\ldots; \quad u_{1,m-1}; u_{2,m-1}; \ldots; u_{n-1,m-1})^T,$$

то система сеточных уравнений (21.13) может быть представлена в обычной для записи СЛАУ форме

$$\mathbf{A}_h \mathbf{u}_h = \mathbf{b}_h, \tag{21.16}$$

где N-мерный вектор $\mathbf{b}_h$ формируется из правых частей уравнений (21.13) в соответствии с тем, как сформирован вектор $\mathbf{u}_h$, а $N \times N$-матрица $\mathbf{A}_h$ — из коэффициентов этих уравнений.

Для решения N-мерной линейной алгебраической системы (21.16), в принципе, можно применять любые методы, прямые и итерационные, в частности, методы, изложению которых посвящены гл.2 и гл.3. Если максимальная размерность N систем (21.16) невелика, то на выборе способа их решения можно не заострять внимания и использовать стандартные методы, например, метод Гаусса. Однако при решении более-менее сложных прикладных стационарных задач приходится пользоваться сетками, имеющими десятки, а то и сотни тысяч узлов, и, как следствие, такое же громадное число неизвестных в системах вида (21.16). В подобных случаях не обойтись без выбора более эффективных методов по требуемым вычислительным затратам, чем обычный метод Гаусса, количество затрачиваемых арифметических операций в котором составляет величину $O(N^3)$. Ясно, что построение таких эффективных методов немыслимо без учета специфики систем (21.16).

Основной особенностью систем (21.16), составленных из разностных уравнений (21.13), присущей им независимо от вида исходного эллиптического уравнения, конфигурации области Ω, типа граничных условий и способа их аппроксимации, является разреженность матрицы системы $\mathbf{A}_h$, т.е. тот факт, что большинство ее элементов — нули. Структура же матрицы $\mathbf{A}_h$, т.е. расположение в ней ненулевых элементов (элементов, не являющихся заведомо нулями), зависит и от вида исходной области Ω, на которой задано решаемое эллиптическое уравнение, и от типа поставленной для него краевой задачи, и от способа упорядочивания неизвестных u_{ij}.

*) Такой способ упорядочивания называют лексикографическим [138].

При выполнении обычных гауссовых преобразований исключения без учета структуры матрицы $\mathbf{A}_h$ многие ее нулевые элементы постепенно заменяются ненулевыми, т.е. при выполнении прямого хода метода Гаусса матрица $\mathbf{A}_h$ теряет свойство разреженности. Хотя в общем случае такой процесс неизбежен, существует стратегия выбора главного элемента, при которой удается достичь компромисса между численной устойчивостью и минимальным заполнением матрицы ненулевыми элементами, что ведет к уменьшению совокупного числа «существенных» (не над нулями) арифметических операций. Одна из таких ***стратегий*** называется ***стратегией*** или ***схемой Марковица*** и составляет основу ***алгоритмов минимальной степени***, с которыми можно ознакомиться, например, с помощью книг [64, 197].

При решении СЛАУ большой размерности прямыми методами систему обычно сначала представляют в блочной форме, стараясь при этом максимально учесть структуру матрицы системы. Такой подход позволяет резко сократить объем вычислений и снизить требования к оперативной памяти вычислительной машины*), поскольку здесь решение одной большой СЛАУ сводится к решению нескольких СЛАУ меньшей размерности, из которых весьма значительная часть — с заведомо нулевыми матрицами и правыми частями.

Чтобы продемонстрировать, что может дать блочный подход к решению сеточных СЛАУ, рассмотрим разностную схему (21.5)–(21.6) для уравнения Пуассона в прямоугольной области с сеткой с квадратными ячейками. При $h_1 = h_2 = h$ разностное уравнение (21.5) имеет наиболее простой вид:

$$u_{i-1,j} + u_{i+1,j} - 4u_{ij} + u_{i,j-1} + u_{i,j+1} = h^2 f_{ij}, \tag{21.17}$$

где $i = 1, 2, \ldots, n-1$, $j = 1, 2, \ldots, m-1$. Согласно (21.6), это уравнение доопределяется условиями

$$u_{0j} = \varphi_{0j}, \quad u_{nj} = \varphi_{nj} \quad \forall\, j \in \{1, 2, \ldots, m-1\}, \tag{21.18}$$

$$u_{i0} = \varphi_{i0}, \quad u_{im} = \varphi_{im} \quad \forall\, i \in \{1, 2, \ldots, n-1\}. \tag{21.19}$$

*) В случае использования системы параллельной обработки информации разбиение на блоки просто необходимо, но рационально учитывать структуру матриц при этом сложнее.

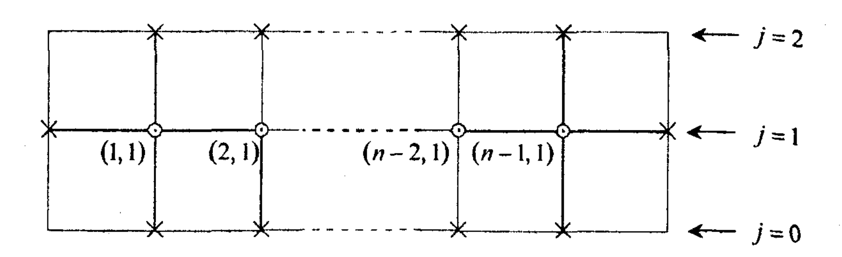


Рис. 21.6. Частный случай сетки рис. 21.1, соответствующий значению $m = 2$

Предположим, что $m = 2$, т.е. сетка Ω состоит только из одной «линейки» внутренних узлов $(i, 1)$ (рис.21.6). Тогда из (21.17) при $j = 1$ получаем уравнение,

$$u_{i-1,1} - 4u_{i1} + u_{i+1,1} = h^2 f_{i1} - u_{i0} - u_{i2},$$

которое с учетом равенств (21.18), (21.19) можно записать в виде системы

$$\begin{cases} -4u_{11} + u_{21} = h^2 f_{11} - \varphi_{10} - \varphi_{12} - \varphi_{01}, \\ u_{i-1,1} - 4u_{i1} + u_{i+1,1} = h^2 f_{i1} - \varphi_{i0} - \varphi_{i2} \quad (i = 2, \ldots, n-2), \\ u_{n-2,1} - 4u_{n-1,1} = h^2 f_{n-1,1} - \varphi_{n-1,0} - \varphi_{n-1,2} - \varphi_{n1}. \end{cases}$$

Эта трехдиагональная $(n-1)$-мерная СЛАУ представляет собой частный случай системы (21.16), которому придаем следующий векторно-матричный вид:

$$\mathbf{T}\mathbf{u}_1 = \widetilde{\mathbf{b}}_1,$$

где

$$\mathbf{T} := \begin{pmatrix} -4 & 1 & 0 & \ldots & 0 & 0 \\ 1 & -4 & 1 & \ldots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & -4 & 1 \\ 0 & 0 & 0 & \ldots & 1 & -4 \end{pmatrix}, \quad \mathbf{u}_1 := \begin{pmatrix} u_{11} \\ u_{21} \\ \ldots \\ u_{n-2,1} \\ u_{n-1,1} \end{pmatrix}, \qquad (21.20)$$

$$\widetilde{b}_1 := \begin{pmatrix} h^2 f_{11} - \varphi_{10} - \varphi_{12} - \varphi_{01} \\ h^2 f_{21} - \varphi_{20} - \varphi_{22} \\ \cdots\cdots\cdots \\ h^2 f_{n-2,1} - \varphi_{n-2,0} - \varphi_{n-2,2} \\ h^2 f_{n-1,1} - \varphi_{n-1,0} - \varphi_{n-1,2} - \varphi_{n1} \end{pmatrix}.$$

Теперь положим $m = 3$, т.е. рассмотрим ситуацию, когда сетка на данном прямоугольнике состоит из двух рядов внутренних узлов (рис. 21.7).

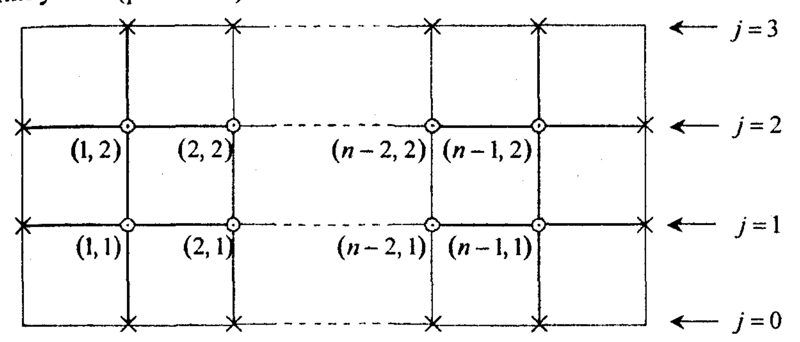


Рис. 21.7. Частный случай сетки рис. 21.1, соответствующий значению $m = 3$

В таком случае, полагая в (21.17) $j = 1$ и $j = 2$, получаем (с учетом условий (21.19)) следующую систему:

$$\begin{cases} u_{i-1,1} - 4u_{i1} + u_{i+1,1} + u_{i2} = h^2 f_{i1} - \varphi_{i0}, \\ u_{i1} + u_{i-1,2} - 4u_{i2} + u_{i+1,2} = h^2 f_{i2} - \varphi_{i3}, \end{cases}$$

где $i = 1, 2, \ldots, n-1$ и известны значения

$$u_{01} = \varphi_{01}, \quad u_{n1} = \varphi_{n1}, \quad u_{02} = \varphi_{02}, \quad u_{n2} = \varphi_{n2}.$$

Если дополнительно к обозначениям (21.20) обозначить

$$\mathbf{u}_2 := \begin{pmatrix} u_{12} \\ u_{22} \\ \ldots \\ u_{n-2,2} \\ u_{n-1,2} \end{pmatrix}, \quad \mathbf{b}_1 := \begin{pmatrix} h^2 f_{11} - \varphi_{10} - \varphi_{01} \\ h^2 f_{21} - \varphi_{20} \\ \cdots\cdots \\ h^2 f_{n-2,1} - \varphi_{n-2,0} \\ h^2 f_{n-1,1} - \varphi_{n-1,0} - \varphi_{n1} \end{pmatrix},$$

$$\widetilde{\mathbf{b}}_2 := \begin{pmatrix} h^2 f_{12} - \varphi_{13} - \varphi_{02} \\ h^2 f_{22} - \varphi_{23} \\ \cdots\cdots\cdots\cdots \\ h^2 f_{n-2,2} - \varphi_{n-2,3} \\ h^2 f_{n-1,2} - \varphi_{n-1,3} - \varphi_{n2} \end{pmatrix},$$

то последнюю систему, как легко убедиться непосредственной проверкой, можно представить в блочном виде

$$\begin{pmatrix} \mathbf{T} & \mathbf{E} \\ \mathbf{E} & \mathbf{T} \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{b}_1 \\ \widetilde{\mathbf{b}}_2 \end{pmatrix},$$

что равносильно записи в виде системы из двух $(n-1)$-мерных векторно-матричных уравнений

$$\begin{cases} \mathbf{T}\mathbf{u}_1 + \mathbf{u}_2 = \mathbf{b}_1, \\ \mathbf{u}_1 + \mathbf{T}\mathbf{u}_2 = \widetilde{\mathbf{b}}_2 \end{cases} \tag{21.21}$$

или одного векторно-матричного уравнения (21.16), где $\mathbf{A}_h$ — симметричная матрица $\begin{pmatrix} \mathbf{T} & \mathbf{E} \\ \mathbf{E} & \mathbf{T} \end{pmatrix}$, а векторы $\mathbf{u}_h$ и $\mathbf{b}_h$ составляются из двух одноименных векторов, именно

$$\mathbf{u}_h = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix}, \quad \mathbf{b}_h = \begin{pmatrix} \mathbf{b}_1 \\ \widetilde{\mathbf{b}}_2 \end{pmatrix}.$$

Если продолжить такой процесс увеличения значений m, то при переходе от $m = 3$ к $m = 4$ система (21.21) заменяется системой вида

$$\begin{cases} \mathbf{T}\mathbf{u}_1 + \mathbf{u}_2 = \mathbf{b}_1, \\ \mathbf{u}_1 + \mathbf{T}\mathbf{u}_2 + \mathbf{u}_3 = \mathbf{b}_2, \\ \qquad \mathbf{u}_2 + \mathbf{T}\mathbf{u}_3 = \widetilde{\mathbf{b}}_3 \end{cases} \tag{21.22}$$

(значок $\sim$ над $\mathbf{b}_k$ показывает, что k-й вектор свободных членов подвержен влиянию граничных условий $u_{1m} = \varphi_{1m}$, $u_{2m} = \varphi_{2m}$), и т.д.

При произвольном $m \geq 4$ соответствующая (21.17)–(21.19) N-мерная система (21.16) имеет блочно-трехдиагональную

структуру

$$\begin{cases} \mathbf{T}\mathbf{u}_1 + \mathbf{u}_2 = \mathbf{b}_1, \\ \mathbf{u}_{j-1} + \mathbf{T}\mathbf{u}_j + \mathbf{u}_{j+1} = \mathbf{b}_j \quad (j = 2, \ldots, m-2), \\ \qquad \mathbf{u}_{m-2} + \mathbf{T}\mathbf{u}_{m-1} = \widetilde{\mathbf{b}}_{m-1}, \end{cases} \tag{21.23}$$

вид которой говорит о возможности обобщить на этот случай известный метод прогонки (2.6). Действительно, такое обобщение существует, причем для блочно-трехдиагональных систем более общего вида

$$\begin{cases} \mathbf{B}_1\mathbf{u}_1 + \mathbf{C}_1\mathbf{u}_2 = \mathbf{F}_1, \\ \mathbf{A}_j\mathbf{u}_{j-1} + \mathbf{B}_j\mathbf{u}_j + \mathbf{C}_j\mathbf{u}_{j+1} = \mathbf{F}_j \quad (j = 2, \ldots, m-2), \\ \qquad \mathbf{A}_{m-1}\mathbf{u}_{m-2} + \mathbf{B}_{m-1}\mathbf{u}_{m-1} = \mathbf{F}_{m-1} \end{cases}$$

(в таком виде можно записывать системы сеточных уравнений (21.13) с пятидиагональными матрицами $\mathbf{A}_h$), и называется ***методом матричной прогонки*** [103, 154, 159, 161]. Хотя его эффективность существенно выше, чем у обычного метода Гаусса, все же он не столь экономичен, как это можно было ожидать, поскольку на каждом этапе прямого хода приходится обращать $(n-1)$-мерные матрицы*).

Имеются и более эффективные прямые методы решения таких специфических СЛАУ, которые возникают при конечноразностной аппроксимации уравнений Пуассона в прямоугольной области, например, ***метод БПФ*** (быстрого преобразования Фурье), опирающийся на метод разделения переменных в системе сеточных уравнений [13, 14, 78, 158]. В частности, для решения систем вида (21.23) разработан весьма экономичный (с числом операций $O(N)$) так называемый ***марш-алгоритм***, описание которого можно найти в книге [153].

21.3. ОБ ИТЕРАЦИОННОМ РЕШЕНИИ СЕТОЧНЫХ УРАВНЕНИЙ

При решении систем сеточных уравнений с большими разреженными матрицами, возникающих при конечноразностной аппроксимации краевых задач для эллиптических уравнений,

---

*) При изменении направления упорядочивания неизвестных (сначала вдоль оси Oy, а затем вдоль оси Ox) нужно обращать $(m-1)$-мерные матрицы. Обычно, упорядочивание производят так, чтобы обращать матрицы меньшей размерности.

часто отдают предпочтение итерационным методам в силу следующих обстоятельств. Во-первых, реализация таких методов намного проще. Это преимущество перед прямыми методами особенно заметно в случаях, когда области Ω, на которых задаются эллиптические уравнения, имеют сложную форму, что порождает проблемы с учетом структуры матриц коэффициентов $\mathbf{A}_h$ в (21.16) при построении прямых алгоритмов исключения с минимальным заполнением и что практически никак не отражается при использовании тех или иных итерационных методов. Во-вторых, итерационные методы заведомо менее требовательны к памяти компьютера, чем любые прямые методы. В-третьих, как правило, использование подходящих итерационных методов позволяет получить выигрыш и в общем объеме вычислений при решении разностным методом с заданной точностью стационарных задач с большим количеством узлов.

Большое удобство при реализации итерационных методов создает тот факт, что представление системы сеточных уравнений в виде (21.16) нужно лишь номинально: для того, чтобы увидеть структуру матрицы $\mathbf{A}_h$ системы, изучить ее свойства, выяснить вопросы применимости того или иного метода, получить оценки погрешности и т. п.; организация же собственно итерационных вычислений производится на основе самой записи разностного уравнения. Отсюда ясно, что здесь нет характерной для прямых методов проблемы возможного замещения нулевых элементов ненулевыми.

Для простоты будем опять ориентироваться на разностную схему (21.17)–(21.19), построенную для задачи (21.1)–(21.2) в прямоугольнике на шаблоне «крест» с шагом h по x и y, имея при этом в виду, что все нижеследующее легко переносится на общий случай сеточных уравнений вида (21.13) (и не только их), а сложности появляются лишь при обосновании сходимости таких обобщений.

Начнем с метода простых итераций.

Учитывая, что при принятом за основу упорядочивании неизвестных u_{ij} по одному из направлений координатных осей матрица $\mathbf{A}_h$ системы (21.16) является пятидиагональной с диагональным преобладанием (в отдельных случаях, где содержится три или четыре ненулевых элемента, — строгим, в остальных — нестрогим, см. (21.23) совместно с (21.20)), естественно попытаться использовать МПИ (3.1) в форме метода Якоби (3.2). Для данного случая, в соответствии с разностным уравнением (21.17), это означает, что основная расчетная формула ***метода Якоби*** есть

$$u_{ij}^{(k+1)} = \frac{1}{4}\left(u_{i-1,j}^{(k)} + u_{i+1,j}^{(k)} + u_{i,j-1}^{(k)} + u_{i,j+1}^{(k)}\right) - \frac{h^2}{4} f_{ij}, \qquad (21.24)$$

где k — номер итерации — принимает последовательно значения $0, 1, 2, \dots$; при каждом значении k, начиная с нулевого (при котором начальные значения $u_{ij}^{(0)}$ неизвестных u_{ij} считаются заданными), индексы i и j изменяются от 1 до $n-1$ и до $m-1$ соответственно, причем порядок их изменения здесь роли не играет. При крайних значениях i и j, т.е. при $i=1$ или $i=n-1$ и $j=1$ или $j=m-1$, хотя бы одна из величин $u_{i-1,j}^{(k)}$, $u_{i+1,j}^{(k)}$, $u_{i,j-1}^{(k)}$, $u_{i,j+1}^{(k)}$ за счет краевых условий (21.18) и (21.19) переходит в разряд известных при любых $k = 0, 1, 2, \dots$.

Наличие нестрогого преобладания в матрице $\mathbf{A}_h$ говорит о том, что ситуация со сходимостью итерационного процесса (21.24) — критическая. Действительно, совершенно очевидно, что естественные легко вычислимые нормы матрицы итерирования в (21.24), подчиненные, например, норме-максимум или норме-сумме вектора $\mathbf{u}_h$, равны единице, т.е. здесь неприменимы теоремы о достаточных условиях сходимости МПИ типа теоремы 3.2. В то же время, более детальное изучение этой матрицы (обозначим ее $\mathbf{J}_h$) показывает [138], что ее спектральный радиус

$$\rho(\mathbf{J}_h) = \cos \pi h \approx 1 - 0.5\pi^2 h^2 < 1 \quad \forall\, h > 0, \qquad (21.25)$$

следовательно, сходимость процесса (21.24) к решению системы сеточных уравнений (21.17) гарантируется теоремой 3.1 о необходимых и достаточных условиях сходимости МПИ. Однако скорость сходимости этого процесса, в силу близости к единице спектрального радиуса матрицы итерирования, будет заведомо низкой и заметно ухудшаться с уменьшением шага, из чего можно сделать вывод о нецелесообразности применения такого метода при сколько-нибудь значительном числе узлов сетки.

Как следует из материала главы 3, в данной ситуации можно надеяться на лучшие результаты (без увеличения объема вычислений), применив к системе уравнений (21.17) ***метод Зейделя***, запись основной расчетной формулы которого, в соответствии с (3.13), имеет вид

$$u_{ij}^{(k+1)} = \frac{1}{4}\left(u_{i-1,j}^{(k+1)} + u_{i+1,j}^{(k)} + u_{i,j-1}^{(k+1)} + u_{i,j+1}^{(k)}\right) - \frac{h^2}{4} f_{ij}, \qquad (21.26)$$

где так же, как и в (21.24), при каждом $k = 0, 1, 2, \dots$ полагаем $i = 1, 2, \dots, n-1$, $j = 1, 2, \dots, m-1$ и учитываем краевые условия (21.18), (21.19). Утверждение о сходимости процесса Зейделя (21.26) можно считать следствием, например, теоремы 3.10.

И метод Якоби (21.24), и метод Зейделя (21.26) (известные в этих специфичных вариантах еще и под названием ***метод Либмана*** [27, 62, 115, 126]) ввиду их медленной сходимости малоприемлемы для расчетов с мелкими сетками, что обусловлено не только большим количеством вычислений, но и повышенным риском накопления при этом значительных погрешностей (см. по этому поводу § 3.7).

Уменьшить потребное для получения каркаса решения с заданной точностью число итераций метода Зейделя (21.26) можно введением в него ***параметра релаксации*** $\omega \in (1, 2)$, т.е. привлекая для решения сеточных уравнений (21.17) метод ПВР (иначе, SOR-метод), описанный в § 3.4*).

В соответствии с записью (3.21) последовательность приближений $\left(u_{ij}^{(k)}\right)$ по методу релаксации к искомым значениям u_{ij} определяется формулой

$$u_{ij}^{(k+1)} = u_{ij}^{(k)} + \omega\left(\hat{u}_{ij}^{(k+1)} - u_{ij}^{(k)}\right), \qquad (21.27)$$

где через $\hat{u}_{ij}^{(k+1)}$ обозначен результат вычислений по формуле Зейделя (21.26), т.е. речь идет о парном, поочередном применении формул (21.26) и (21.27): основной шаг, затем шаг ускорения. Ускорение здесь действительно может быть ощутимым, если сделан подходящий выбор ускоряющего множителя ω. Для каждой конкретной матрицы $\mathbf{A}_h$ системы (21.16) имеется свое оптимальное значение $\omega_{\text{опт}}$, при котором с помощью (21.27) достигается максимальное ускорение итерационного процесса Зейделя. В некоторых случаях это оптимальное значение параметра релаксации может быть заранее подсчитано.

Так, для матриц $\mathbf{A}_h$, «упорядоченно согласованных со свойством A» (подробно об этом см. в [27]), установлена связь между собственными числами соответствующих им матриц итерирования $\mathbf{J}_h$ метода Якоби и матриц итерирования $\mathbf{S}_h(\omega)$ метода ПВР (SOR-метода), записанного в эквивалентной МПИ форме, что позволяет подобрать значение $\omega = \omega_{\text{опт}}$ таким, при котором минимизируется спектральный радиус $\rho(\mathbf{S}_h)$ и обеспечивается наиболее быстрая сходимость последовательности приближений, получаемых методом ПВР.

*) Применительно к рассматриваемой задаче этот метод иногда называют ***ускоренным методом Либмана***, а параметр ω — ***ускоряющим множителем*** [115].

В частности, матрица $\mathbf{A}_h$ системы сеточных уравнений (21.17) упомянутым свойством обладает, и с учетом выражения (21.25) по формуле [27, 138]

$$\omega_{\text{опт}} = \frac{2}{1 + \sqrt{1 - \rho^2(\mathbf{J}_h)}} \qquad (21.28)$$

для релаксационного процесса (21.26)–(21.27) находим

$$\omega_{\text{опт}} = \frac{2}{1 + \sqrt{1 - \cos^2 \pi h}} = \frac{2}{1 + \sin \pi h} \approx \frac{2}{1 + \pi h}. \qquad (21.29)$$

Выигрыш в скорости сходимости при этом нетрудно оценить, зная, что спектральный радиус матрицы $\mathbf{S}_h(\omega_{\text{опт}})$ для систем с охарактеризованными выше матрицами $\mathbf{A}_h$ на единицу отличается от параметра $\omega_{\text{опт}} \in (1, 2)$:

$$\rho(\mathbf{S}_h(\omega_{\text{опт}})) = \omega_{\text{опт}} - 1. \qquad (21.30)$$

Подсчитав величины $-1/\ln(\rho(\mathbf{S}_h(\omega_{\text{опт}}))$, $-1/\ln(\rho(\mathbf{S}_h(1))$ и $-1/\ln(\rho(\mathbf{J}_n))$, показывающие количество итераций, затрачиваемых рассматриваемыми процессами (соответственно, ПВР, Зейделя и Якоби) на установление в результате одного верного знака во всех компонентах, можно судить о сравнительном преимуществе метода ПВР, быстро возрастающем с уменьшением шага h (см. упр. 21.5).

Для матриц $\mathbf{A}_h$, не обладающих нужными свойствами, применение формулы (21.28) (а также (21.30)) не оправдано, и значение параметра ω, обеспечивающее ускорение итерационного процесса, тем или иным способом подбирают. Это часто делают и в случаях, когда формально есть основания вычислить $\omega_{\text{опт}}$ по формуле (21.28), но значение $\rho(\mathbf{J}_n)$ неизвестно.

Замечание 21.1. Существует несколько модификаций метода ПВР. Укажем на две из них [137]. Одна модификация опирается на запись системы сеточных уравнений в блочном виде, поблочную реализацию метода Зейделя (Гаусса–Зейделя) и повекторное ускорение по формуле, подобной формуле (21.27); такой метод называют ***блочным методом ПВР (SOR)***. Другая модификация направлена на устранение неравноправия неизвестных, присущего методу Зейделя (для него небезразличен порядок подключения неизвестных при формировании вектора $\mathbf{u}_h$ в системе (21.16)). Сначала строится ***симметричный метод Зейделя***, состоящий из двух полушагов, на каждом из которых неизвестные упорядочиваются поразному (в прямом и в обратном направлениях, что означает перемену ролями нижней и верхней треугольных матриц при аддитивном разбиении исходной матрицы системы), затем в оба этих полушага обычным образом

вводится параметр релаксации; в итоге получается метод, называемый ***методом симметричной последовательной верхней релаксации (SSOR)***. Утверждение о его сходимости не отличается от теоремы 3.12.

Пример 21.1. Рассмотрим применение методов Зейделя и ПВР к нахождению таблицы значений функции $\psi(x, y)$ из уравнения

$$\frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}-\frac{3}{x}\frac{\partial\psi}{\partial x}+2=0, \tag{21.31}$$

определенного в прямоугольнике $\Omega:=[R-b,\, R+b]\times[-a,\, a]$, с граничным условием

$$\psi(x, y)\big|_\Gamma=0. \tag{21.32}$$

Такая таблица может служить картой напряжений в прямоугольном сечении нагруженной цилиндрической пружины (рис. 21.8) и использоваться в конструкторских расчетах при вычислении коэффициентов перенапряжений (обычно задаваемых номограммами [22]).

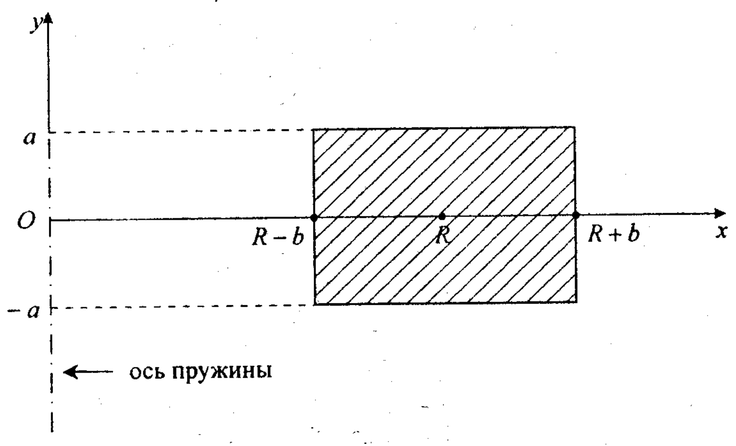


Рис. 21.8. Сечение одного витка цилиндрической пружины

Введя на Ω сетку с шагами $h_1=\dfrac{2b}{n}$ по x (в соответствии с чем $x_i=R-b+ih_1$) и $h_2=\dfrac{2a}{m}$ по y и применив симметричные аппроксимации

$$\left.\frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}\right|_{\substack{x=x_i\\ y=y_j}}\approx\frac{\psi_{i-1,j}-2\psi_{ij}+\psi_{i+1,j}}{h_1^2}+\frac{\psi_{i,j-1}-2\psi_{ij}+\psi_{i,j+1}}{h_2^2},$$

$$\left.\frac{\partial\psi}{\partial x}\right|_{\substack{x=x_i\\ y=y_j}}\approx\frac{\psi_{i+1,j}-\psi_{i-1,j}}{2h_1},$$

данное эллиптическое уравнение (21.31) аппроксимируем системой сеточных уравнений (типа 21.13)

$$A\psi_{ij}=B_i\psi_{i-1,j}+C_i\psi_{i+1,j}+D\psi_{i,j-1}+E\psi_{i,j+1}+2, \tag{21.33}$$

где:

$$A:=\frac{2}{h_1^2}+\frac{2}{h_2^2},\quad B_i:=\frac{1}{h_1^2}+\frac{3}{2h_1(R-b+ih_1)},\quad C_i:=\frac{1}{h_1^2}-\frac{3}{2h_1(R-b+ih_1)},$$

$$D:=\frac{1}{h_2^2},\quad E:=\frac{1}{h_2^2};\quad i=1,2,\ldots,n-1,\quad j=1,2,\ldots,m-1.$$

Дополнительными к (21.33) служат равенства

$$\psi_{0j}=\psi_{nj}=\psi_{i0}=\psi_{im}=0,$$

соответствующие заданному граничному условию (21.32).

Решение СЛАУ (21.33) методом Зейделя означает проведение итераций по формуле

$$\psi_{ij}^{(k+1)}=\frac{1}{A}\left(B_i\psi_{i-1,j}^{(k+1)}+C_i\psi_{i+1,j}^{(k)}+D\psi_{i,j-1}^{(k+1)}+E\psi_{i,j+1}^{(k)}+2\right), \tag{21.34}$$

начинаемых с задаваемых значений $\psi_{ij}^{(0)}$ и заканчиваемых по выполнении некоторого критерия останова. Превращение метода Зейделя в метод ПВР сводится к тому, что при введенном предварительно значении параметра ω, выполнив вычисление $\psi_{ij}^{(k+1)}$ по формуле (21.34), делаем присвоение

$$\hat\psi_{ij}^{(k+1)}:=\psi_{ij}^{(k+1)}$$

и вычисляем новое (уточненное) значение

$$\psi_{ij}^{(k+1)}=\psi_{ij}^{(k)}+\omega\left(\hat\psi_{ij}^{(k+1)}-\psi_{ij}^{(k)}\right)$$

(все это при $k=0,1,2,\ldots,\quad i=1,2,\ldots,n-1,\quad j=1,2,\ldots,m-1$).

Вычисления по данным формулам при фиксированных значениях

$$R:=18,\quad a:=3,\quad b:=6,\quad n:=16,\quad m:=8,\quad \omega:=1.539,$$

начальном приближении $\psi_{ij}^{(0)}:=0$ и критерии окончания итераций

$$\sqrt{\sum_{i,j}\left(\psi_{ij}^{(k+1)}-\psi_{ij}^{(k)}\right)^2}\le 0.0001$$

дают следующую карту напряжений во внутренних узлах сетки, определенной указанными значениями постоянных (приводимая таблица значений ψ_{ij} вдвое прорежена по направлению x по сравнению с расчетной):

0.534	0.763	0.866	0.902	0.887	0.804	0.584
0.868	1.279	1.468	1.535	1.508	1.356	0.959
1.052	1.578	1.823	1.910	1.875	1.678	1.170
1.111	1.675	1.940	2.034	1.996	1.783	1.238
1.052	1.578	1.823	1.910	1.875	1.678	1.170
0.868	1.279	1.468	1.535	1.508	1.356	0.959
0.534	0.763	0.866	0.902	0.887	0.804	0.584 .

Число итераций, выполненных методом Зейделя и методом ПВР, — 98 и 23 соответственно.

21.4. МЕТОДЫ УСТАНОВЛЕНИЯ

Знакомство с основной идеей методов установления на абстрактном уровне, т.е. для произвольных СЛАУ безотносительно их происхождения, у нас состоялось в § 3.5. В развитие этой идеи там же представлена общая концепция двухслойных итерационных методов, к которым, кстати, можно отнести и рассмотренные в предыдущем параграфе методы Якоби, Зейделя, ПВР.

Конкретизация решаемых стационарных задач и учет их физической интерпретации позволяет придать идее установления определенный содержательный смысл и подобрать наиболее эффективные методы решения систем сеточных уравнений среди множества двухслойных итерационных методов.

Поставим рядом рассматриваемое в этой главе уравнение Пуассона

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=f(x,y) \qquad (21.35)$$

и двумерное уравнение теплопроводности

$$\frac{\partial u}{\partial t}=a^2\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right)+f(x,y,t), \qquad (21.36)$$

численному решению которого посвящен § 20.6.

В частном случае, когда в (21.36) $a^2=1$ и функция f не зависит от переменной t, т.е. когда уравнение (21.36) берется в виде

$$\frac{\partial u}{\partial t}=\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}-f(x,y), \qquad (21.37)$$

при $\frac{\partial u}{\partial t}\equiv 0$ оно совпадает с уравнением Пуассона (21.35). Если для уравнения теплопроводности (21.37) задать не зависящее от времени t граничное условие

$$u|_{\Gamma}=\varphi(x,y), \qquad (21.38)$$

соответствующее условию (21.2) в рассматриваемой задаче Дирихле, то при любых начальных условиях на функцию $u(x,y,t)$ в области Ω с границей Γ

$$u(x,y,0)=u^0(x,y) \qquad (21.39)$$

их влияние на решение $u(x,y,t)$ с течением времени t ослабевает и при $t\to\infty$

$$u(x,y,t)\to u(x,y), \qquad \frac{\partial u}{\partial t}\to 0.$$

Таким образом, стационарную краевую задачу (21.35), (21.38) можно считать предельной для эволюционной начально-граничной задачи (21.37), (21.38), (21.39) и получать ее приближенное решение $u(x,y)$ в Ω, находя решение $u(x,y,t)$ последней для достаточно больших t. Поскольку применение разностного метода для получения каркаса решения нестационарной задачи (21.37)–(21.39) означает последовательные, от t_k к t_{k+1} ($k=0,1,\ldots$), вычисления, процесс таких вычислений должен продолжаться до тех пор, пока не произойдет выход на стационарный режим, т.е. значения решения на $(k+1)$-м слое будут практически (в пределах заданной точности) совпадать с соответствующими значениями k-го слоя во всех внутренних узлах сетки. Это и есть пошаговое установление вычислительного процесса, соответствующее выходу физического эволюционного процесса в стационарный, установившийся режим. Учитывая экспоненциальный характер затухания влияния начального условия (21.39), такое установление должно происходить достаточно быстро, и, очевидно, тем быстрее, чем лучше начальная функция $u^0(x,y)$ в (21.39) отражает искомое стационарное распределение температуры, задаваемое уравнением (21.35) и граничным условием (21.38).

Зная, что одним из эффективнейших методов решения задач теплопроводности с двумя пространственными переменными является метод переменных направлений (Писмэна–Рэчфорда), запишем соответствующие этому двухслойному методу формулы применительно к рассматриваемой задаче Дирихле для уравнения Пуассона*). Главной особенностью, отличающей применение метода переменных направлений к стационарным задачам, является то, что здесь шаги по времени t, которое в записи стационарного уравнения явно отсутствует, следует расценивать как итерационные шаги, т.е. индекс k, означавший в § 20.6 номер слоя пространственно-временной сетки, в нижеследующих формулах интерпретируется как номер итерации (что будем подчеркивать, заключая его в скобки). В связи с этим к такому методу применяют название ***итерационный метод переменных направлений***.

---

*) Иногда при решении стационарных задач используют и трехслойные итерационные методы, сводя эти задачи к эволюционным задачам с гиперболическими уравнениями.

В соответствии с формулами (20.60), (20.61) для задачи (21.35), (21.38) на введенной ранее на Ω сетке узлов $(x_i; y_j)$ имеем:

$$u_{i-1,j}^{\left(k+\frac{1}{2}\right)} - \left(2+\frac{2h^2}{\tau}\right)u_{ij}^{\left(k+\frac{1}{2}\right)} + u_{i+1,j}^{\left(k+\frac{1}{2}\right)} = h^2 f_{ij} - u_{i,j-1}^{(k)} + \left(2-\frac{2h^2}{\tau}\right)u_{ij}^{(k)} - u_{i,j+1}^{(k)}, \quad (21.40)$$

$$u_{i,j-1}^{(k+1)} - \left(2+\frac{2h^2}{\tau}\right)u_{ij}^{(k+1)} + u_{i,j+1}^{(k+1)} = h^2 f_{ij} - u_{i-1,j}^{\left(k+\frac{1}{2}\right)} + \left(2-\frac{2h^2}{\tau}\right)u_{ij}^{\left(k+\frac{1}{2}\right)} - u_{i+1,j}^{\left(k+\frac{1}{2}\right)}. \quad (21.41)$$

Определяемый этими формулами ***продольно-поперечный метод*** можно назвать полуитерационным методом в двух смыслах. Во-первых, здесь делается как бы половина итерационного шага одного направления, затем — вторая половина шага вдоль перпендикулярного прежнему направления. Во-вторых, для осуществления каждой из этих полуитераций нужно решать СЛАУ с трехдиагональной матрицей, для чего применяют прямой метод, например, метод прогонки.

Для начала итераций в формуле (21.40) полагаем $k=0$, подставляем соответствующие выбранной начальной функции $u^0(x, y)$ значения*)

$$u_{ij}^{(0)} = u^0(x_i, y_j),$$

фиксируем $j = 1, \ldots, m-1$ и при каждом фиксированном j изменяем i от 1 до $n-1$; получаемые при этом трехточечные разностные уравнения второго порядка дополняем краевыми условиями в соответствии с вытекающими из (21.38) равенствами

$$u_{0j}^{(k)} = u_{0j}^{\left(k+\frac{1}{2}\right)} = \varphi(x_0, y_j) \quad \forall j \in \{1, \ldots, m-1\} \quad \forall k \in \mathbf{N}_0$$

и решаем прогонкой. Пользуясь найденными значениями $u_{ij}^{\left(\frac{1}{2}\right)}$ при том же k аналогично поступаем с формулой (21.41): фиксируем поочередно i от 1 до $n-1$ и при каждом фиксированном i, изменяя j от 1 до $m-1$, с помощью дополнительных равенств

$$u_{i0}^{(k)} = u_{i0}^{\left(k+\frac{1}{2}\right)} = \varphi(x_i, y_0) \quad \forall i \in \{1, \ldots, n-1\} \quad \forall k \in \mathbf{N}_0$$

прогонкой вычисляем значения $u_{ij}^{(1)}$, служащие результатом первого полного итерационного шага. Затем переходим к следующей итерации и т. д. Окончание этого процесса может быть, например, таким:

$$\max_{i,\,j}\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right| \le \varepsilon \quad \Rightarrow \quad u_{ij} :\approx u_{ij}^{(k+1)}.$$

Другим критерием окончания итераций по формулам (21.40)–(21.41) можно считать условие

$$\frac{\left|\| u_{ij}^{(k+1)} \| - \| u_{ij}^{(k)} \|\right|}{\tau} \le \varepsilon,$$

означающее факт установления процесса с точностью ε: $\left|\frac{\partial u}{\partial t}\right| \le \varepsilon$ [92].

Отдельный разговор о том, как распорядиться ***итерационным параметром*** τ в рассматриваемом методе, выбор которого, несомненно, отражается на скорости сходимости метода. Вопрос этот весьма непрост, и нельзя сказать, что на него всегда можно дать исчерпывающий ответ, устраивающий практиков.

Ясно, что параметр τ может быть как одним и тем же постоянным в обеих формулах вида (21.40) и (21.41), так и различным: в одном направлении — τ_1, а в другом — τ_2; в любом случае, должны существовать оптимальные значения $\tau_{\text{опт}}$ (для каждой эллиптической задачи, для каждой области Ω, для каждой сетки свои), при которых каркас решения с нужной точностью находится за наименьшее число итерационных шагов. Поиск таких значений $\tau_{\text{опт}}$ производится с помощью изучения погрешности метода, т.е. разности между каркасом точного решения на используемой сетке $\Omega_{h_1 h_2}$ и результатом k-й итерации данным методом. Имеются некоторые общие утверждения о том, какими должны быть значения $\tau_{\text{опт}}$ в зависимости от свойств соответствующих задаче и методу разностных операторов [158, 159], и конкретные рекомендации в случаях, когда оговоренные свойства этих операторов изучены и известны, например, их собственные значения или хотя бы границы спектра [78, 158, 159].

В частности, в методе, определяемом формулами (21.40), (21.41), для уравнения Пуассона (21.35) в прямоугольнике Ω таким оптимальным значением параметра τ считается значение

$$\tau_{\text{опт}} = \frac{h^2}{2\sin(\pi h)} \approx \frac{h}{2\pi},$$

обеспечивающее достижение точности ε при решении сеточных уравнений приблизительно за $\left[\frac{-\ln\varepsilon}{2\pi h}\right]$ число итераций.

*) Если нет хорошего начального приближения $\left(u_{ij}^{(0)}\right)$, здесь особенно кстати может оказаться использование сгущающихся сеток: получая «дешевое» решение на крупной сетке, принимаем его расширение на более мелкоячеистую сетку в качестве начального приближения для новых итераций, и т.д.

Большой выигрыш в числе итераций может быть получен варьированием значений итерационного параметра от итерации к итерации. В некоторых случаях удается выписать конечные наборы таких значений τ_k (связываемые обычно с корнями многочленов Чебышева и задаваемые после того, когда уже известно требуемое для достижения нужной точности число итерационных шагов), которые обеспечивают наиболее быстрое затухание начальных данных в методах установления.

В оптимизации значений итерационных параметров τ или их конечных последовательностей (τ_k) нуждаются также так называемые ***попеременно-треугольные итерационные методы*** (ПТИМ), относящиеся к семейству неявных двухслойных методов вида (3.28) и на модельных задачах типа задачи Дирихле для уравнения Пуассона в прямоугольной области демонстрирующие самую высокую эффективность. Согласно [159], асимптотическая (при $h \to 0$) оценка числа итераций, позволяющего найти каркас приближенного решения такой задачи с точностью ε, составляет для ПТИМ с оптимальным (чебышевским) набором параметров τ_k величину порядка $\frac{1}{2\sqrt{\pi h}}\ln\frac{1}{\varepsilon}$, а с одним фиксированным значением $\tau = \tau_{\text{опт}}$ — порядка $\frac{1}{2\pi h}\ln\frac{1}{\varepsilon}$ (как и у итерационного метода переменных направлений). Для применимости попеременно-треугольного метода к системе (21.16) сеточных уравнений требуется, чтобы ее матрица $\mathbf{A}_h$ была симметричной положительно определенной.

УПРАЖНЕНИЯ

21.1. Составьте систему сеточных уравнений, аппроксимирующих на квадратной сетке с шагом $h = 0.5$ уравнение

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = xy,$$

заданное в полукруге $\begin{cases} x^2 + y^2 < 1, \\ y > 0 \end{cases}$ с границей $\mathbf{\Gamma}$, в случае граничных условий вида:

а) $u(x, y)\big|_{(x;y)\in\Gamma} = |x + y| - 1$;

б) $\left[u - \frac{\partial u}{\partial \mathbf{n}}\right]_{(x;y)\in\Gamma} = \begin{cases} y - \sin \pi x, & y > 0, \\ 0, & y = 0. \end{cases}$

Каковы порядки аппроксимации построенных разностных схем?

Найдите решения полученных СЛАУ.

21.2. Запишите выражения векторов $\mathbf{b}_2$, $\tilde{\mathbf{b}}_3$ в правой части системы (21.22) и вектора $\mathbf{b}_j$ в правой части блочно-трехдиагональной СЛАУ (21.23) при $j \in \{2, 3, \ldots, m-2\}$. В каком виде следует представить вектор $\mathbf{b}_j$, чтобы его выражение удовлетворяло также значениям $j = 1$ и $j = m - 1$? [138].

21.3. По аналогии с выводом метода прогонки для решения трехдиагональных СЛАУ (§ 2.6) выведите расчетные формулы (на блочном уровне) метода матричной прогонки для решения блочно-трехдиагональных СЛАУ, конкретно, систем сеточных уравнений (21.17), соответствующих задаче Дирихле для уравнения Пуассона (21.1)–(21.2), записанных в виде (21.23).

21.4. Запишите метод ПВР (21.26)–(21.27) для решения сеточных уравнений (21.17) в виде одной расчетной формулы (3.23) [161].

21.5. Пользуясь формулами (21.25), (21.29) и (21.30), сравните выигрыш в числе итераций при решении систем сеточных уравнений (21.17) методом Якоби (21.24) и методом ПВР (21.27) с $\omega = \omega_{\text{опт}}$ из (21.28) при:

а) $h = 0.1$; б) $h = 0.01$; в) $h = 0.001$.

21.6. А) Перепишите расчетные формулы (21.40)–(21.41) итерационного метода переменных направлений для случая, когда шаги сетки по координатам неодинаковы.

Б) Обобщите формулы (21.40)–(21.41) на случай первой краевой задачи для произвольного двумерного уравнения эллиптического типа (21.7).

21.7. Примените итерационный метод переменных направлений к граничной задаче (21.31)–(21.32) примера 21.1 (см. § 21.3) в условиях, при которых представлены в нем численные результаты. Варьируя значения итерационного параметра τ, попытайтесь минимизировать количество итераций.

ЗАКЛЮЧИТЕЛЬНОЕ ЗАМЕЧАНИЕ

(*о том, чего в этой книге нет, но автору хотелось бы видеть*)

Несмотря на внушительный объем данной книги, имеется ряд вопросов, которые тесно связаны с материалом книги, но по разным причинам либо вовсе в ней не затрагивались, либо рассматривались лишь поверхностно без погружения в соответствующую теорию, либо только упоминались в том или ином контексте. Чтобы читатель, изучающий или применяющий вычислительную математику, не упускал эти вопросы и темы из виду, перечислим их, сопроводив ссылками на учебные пособия и монографии, с помощью которых можно хотя бы частично восполнить соответствующие пробелы. Перечень этот приводится в порядке условно первого соприкосновения с текстом книги (без учета того, идет ли речь о частном вопросе или о крупном разделе) и не претендует на полноту.

1. Сообщение сведений о мультипроцессорных системах обработки информации и о сравнительных возможностях распараллеливания и конвейеризации известных методов вычислительной линейной алгебры [2, 29, 137, 146].

2. Вывод формул, обоснование и использование сингулярных разложений матриц [43, 99, 135].

3. Более детальное изложение процесса QR-факторизации матриц и QR-алгоритма решения полной алгебраической проблемы собственных значений [42, 43, 54, 71, 75, 137, 141, 179].

4. Подробное описание методов локализации (как вещественных, так и комплексных) корней многочленов с вещественными коэффициентами и метода Лобачевского–Греффе их вычисления [17, 20, 61, 72].

5. Краткое изложение основных численных методов решения задач минимизации функции одной и нескольких переменных [3, 32, 49, 52, 68, 73, 78, 108, 139, 178].

6. Введение понятия чебышевской интерполяции и рассмотрение алгоритмов Ремеза и Валле–Пуссена построения многочленов, близких к многочленам наилучшего равномерного приближения [19, 63, 73, 107, 134, 150].

7. Демонстрация возможностей и достоинств использования для аппроксимации функций систем из тригонометрических и рациональных функций, описание и простейшие применения быстрого дискретного преобразования Фурье, ознакомление с идеей функций-всплесков [14, 16, 18, 82, 88, 159, 161].

8. Показ техники параметрического представления многочленов от одного и двух переменных и ее использования при построении кривых и поверхностей Безье и Кастельжо [82, 119, 120].

9. Ознакомление с методикой использования R-функций Рвачева [143].

10. Исследование погрешности построенных в гл. 11 сплайнов и построение сплайнов периодических, сглаживающих и двумерных [1, 24, 91, 111, 114, 167].

11. Охват многомерного случая при рассмотрении задач интерполирования и численного интегрирования. Описание метода статистических испытаний (Монте-Карло) вычисления кратных интегралов [12, 13, 14, 19, 61, 70, 71, 78, 104, 122, 123, 143, 178].

12. Более глубокое изучение одношаговых многостадийных методов Рунге-Кутты решения задачи Коши для ОДУ и базирующихся на нем алгоритмов [6, 20, 60, 185].

13. Построение регуляризирующих алгоритмов для численного решения интегральных уравнений Фредгольма первого рода с описанием способов выбора параметра регуляризации [11, 13, 14, 28, 34, 41, 73, 78, 128, 175].

14. Более строгое изложение сеточных методов (в частности, проекционной природы), позволяющих находить обобщенные решения задач математической физики, сопровождаемое введением пространств Соболева [9, 117, 136, 155, 160, 172, 191].

15. Конкретизация определений устойчивости для разностных схем, аппроксимирующих стационарные и эволюционные задачи математической физики, и проведение соответствующих им исследований на устойчивость [13, 14, 20, 55, 103, 154–159, 181].

16. Рассмотрение идеи расщепления, применяемой при решении систем сеточных уравнений, на матрично-операторном уровне [43, 117, 118, 153].

17. Ознакомление с методом фиктивных областей (погружение произвольной конечной плоской области задания двумерной краевой задачи в прямоугольник) [14, 117] и с многосеточным методом Федоренко, позволяющим эффективно использовать чередование решения сеточных уравнений на грубых и мелких сетках [14, 56, 152, 181, 191].

18. Более подробное рассмотрение блочного подхода к решению систем сеточных уравнений, порождаемых конечноразностными и конечноэлементными методами решения стационарных задач математической физики [29, 64, 137].

19. Описание (хотя бы на идейном уровне) методов граничных интегральных уравнений и граничных элементов решения краевых задач математической физики [152, 182].

20. Повсеместный подсчет количества требуемых для реализации численных методов арифметических операций или его асимптотики [14, 78, 159].

***ПРИЛОЖЕНИЕ* 1**

НЕКОТОРЫЕ СВЕДЕНИЯ ИЗ ФУНКЦИОНАЛЬНОГО АНАЛИЗА

1. МЕТРИЧЕСКИЕ ПРОСТРАНСТВА

Пусть X — множество элементов произвольной природы, объединенных по какому-либо признаку.

Определение 1. *Множество X называется **метрическим пространством**, если любой паре его элементов x, y сопоставляется вещественное число $\rho(x, y)$, называемое **расстоянием** между x и y или **метрикой** и удовлетворяющее следующим трем аксиомам (неотрицательности, симметричности и треугольника):*

1) $\rho(x, y) \geq 0$, *причем* $\rho(x, y) = 0 \iff x = y$;
2) $\rho(x, y) = \rho(y, x)$;
3) $\rho(x, y) \leq \rho(x, z) + \rho(z, y) \quad \forall z \in X$.

Например, множества $\mathbf{Q}$, $\mathbf{R}$, $\mathbf{C}$ рациональных, вещественных и комплексных чисел можно считать метрическими пространствами с естественной для числовых множеств метрикой $\rho(x, y) = |x - y|$.

Аксиомы 1)–3) определяют метрику неоднозначно, а следовательно, на одном и том же множестве X могут быть заданы разные метрические пространства, т.е. метрическим пространством, вообще говоря, считается пара (X, ρ). Однако часто метрическое пространство обозначают одной буквой, совпадающей с обозначением исходного множества, когда безразлично, о какой метрике идет речь, или присваивают специальные наименования наиболее употребительным конкретным метрическим пространствам.

На метрические пространства распространяются многие основные понятия классического анализа. Приведем некоторые из них.

Определение 2. *Элемент x метрического пространства X называется **пределом** бесконечной последовательности элементов x_n из X, если $\rho(x_n, x) \xrightarrow[n\to\infty]{} 0$. В этом случае говорят, что последовательность (x_n) **сходится** к x по метрике пространства X, и отражают это записью $x_n \to x$ или $\lim\limits_{n\to\infty} x_n = x$.*

Единственность предела выводится из аксиомы треугольника.

Определение 3. *Последовательность (x_n) элементов метрического пространства X называется **фундаментальной** (или сходящейся в себе), если $\rho(x_n, x_m) \to 0$ при $n, m \to \infty$.*

Легко видеть, что всякая фундаментальная последовательность ограничена, т.е.

$$\forall k \in \mathbf{N} \quad \exists r > 0 \quad \forall n \in \mathbf{N}: \quad \rho(x_n, x_k) < r.$$

Фундаментальность последовательности элементов $x_n \in X$ является необходимым условием для ее сходимости. Действительно, если $x = \lim x_n$, то $\rho(x_n, x) \to 0$ и $\rho(x_m, x) \to 0$; но тогда по аксиоме треугольника следует

$$\rho(x_n, x_m) \leq \rho(x_n, x) + \rho(x, x_m) \xrightarrow[\substack{n\to\infty \\ m\to\infty}]{} 0,$$

т.е. фундаментальность (x_n).

Обратное неверно: не всякая фундаментальная последовательность элементов метрического пространства имеет предел. Например, в упомянутом пространстве $\mathbf{Q}$ рациональных чисел с метрикой $\rho(x, y) = |x - y|$ можно построить сколько угодно фундаментальных последовательностей, сходящихся к иррациональным числам (пределы есть, но они принадлежат не этому пространству, а его расширению).

Определение 4. *Метрическое пространство X называется **полным**, если в нем всякая фундаментальная последовательность имеет предел.*

Возвращаясь к примеру, отметим, что метрическое пространство $\mathbf{Q}$ не обладает полнотой, а множества $\mathbf{R}$ вещественных чисел и $\mathbf{C}$ комплексных чисел — полные метрические пространства.

Известно, что к а ж д о е вещественное число представимо как предел последовательности рациональных чисел. Следовательно, если метрическое пространство $\mathbf{Q}$ пополнить его всевозможными предельными точками, то получится метрическое пространство $\mathbf{R}$, т.е. $\mathbf{Q}$ всюду плотно в $\mathbf{R}$ согласно следующему определению.

Определение 5. *Множество M элементов метрического пространства X называется **всюду плотным** в X, если каждый элемент x из X представим в виде предела последовательности (x_n) элементов из M.*

Определение 6. *Метрическое пространство X называется **сепарабельным**, если в нем существует счетное (или, в частности, конечное) всюду плотное подмножество.*

Так как множество рациональных чисел $\mathbf{Q}$ — счетное и $\mathbf{Q}$ всюду плотно в $\mathbf{R}$, то метрическое пространство $\mathbf{R}$ вещественных чисел сепарабельно.

Неоднозначность построения метрик по определяющей их системе аксиом для одного и того же множества X требует установления отношения эквивалентности между ними.

Определение 7. *Две метрики, определенные на одном и том же множестве X, называются **эквивалентными**, если сходимость последовательности элементов из X по одной их них означает сходимость этой последовательности и по другой метрике.*

Равносильным данному является другое определение эквивалентности метрик $\rho_1(x, y)$ и $\rho_2(x, y)$ на множестве X: через существование чисел $c_1 > 0$, $c_2 > 0$ таких, что

$$c_1 \rho_1(x, y) \leq \rho_2(x, y) \leq c_2 \rho_1(x, y) \quad \forall x, y \in X.$$

Примеры метрических пространств. 1) На множестве $\mathbf{R}_n$, элементами которого служат упорядоченные совокупности из n вещественных чисел (называемые n-мерными векторами, кортежами или энками) $\mathbf{x}=(x_1,\ldots,x_n)$, $\mathbf{y}=(y_1,\ldots,y_n)$, ..., наиболее употребительными являются следующие метрики:

а) $\rho(\mathbf{x},\mathbf{y})=\max\limits_{i\in\{1,\ldots,n\}}|x_i-y_i|$;

б) $\rho(\mathbf{x},\mathbf{y})=\sqrt{\sum\limits_{i=1}^{n}(x_i-y_i)^2}$ (*евклидова метрика*);

в) $\rho(\mathbf{x},\mathbf{y})=\sum\limits_{i=1}^{n}|x_i-y_i|$.

Эти разные метрики приводят к различным уточнениям понятия сходимости в определяемых ими на $\mathbf{R}_n$ метрических пространствах, однако доказано, что *в конечномерных пространствах все метрики эквивалентны*.

Аналогично метризуется и множество $\mathbf{C}_n$ векторов, координатами которых являются комплексные числа.

2) Рассмотрим множество, элементами которого служат бесконечные последовательности вещественных чисел: $\mathbf{x}=(x_1,x_2,\ldots,x_n,\ldots)$, $\mathbf{y}=(y_1,y_2,\ldots,y_n,\ldots)$, ..., причем такие, для которых имеет место сходимость рядов из квадратов этих чисел, т.е. $\sum\limits_{i=1}^{\infty}x_i^2<\infty$, $\sum\limits_{i=1}^{\infty}y_i^2<\infty$, ... Введя на этом множестве метрику

$$\rho(\mathbf{x},\mathbf{y})=\sqrt{\sum_{i=1}^{\infty}(x_i-y_i)^2}\,,$$

приходим к полному сепарабельному метрическому пространству, обозначаемому l_2 или l^2. Через метрику

$$\rho(\mathbf{x},\mathbf{y})=\left(\sum_{i=1}^{\infty}(x_i-y_i)^p\right)^{\frac{1}{p}} \qquad (1)$$

пространство l_2 обобщается к пространству l_p (или l^p), где $p\geq 1$.

3) Пусть X — множество всех вещественнозначных функций $x(t)$, $y(t)$, ..., определенных и непрерывных при $t\in[a,b]$. Введя метрику

$$\rho(x,y)=\max_{t\in[a,b]}|x(t)-y(t)|,$$

получаем метрическое пространство, называемое ***пространством непрерывных функций*** и обозначаемое $C[a,b]$, $C([a,b])$ или $C_{[a,b]}$. В отличие от этой метрики, называемой ***равномерной*** или ***чебышевской***, на том же множестве непрерывных на $[a,b]$ функций метрику задают также равенствами

$$\rho(x,y)=\sqrt{\int_a^b(x(t)-y(t))^2\,dt} \quad \text{и} \quad \rho(x,y)=\int_a^b|x(t)-y(t)|dt,$$

приходя к метрическим пространствам, обозначаемым как $C_2[a,b]$ или $C_{L_2}[a,b]$ и $C_1([a,b])$ или $C_L[a,b]$ соответственно. На этом множестве уже нет эквивалентности метрик: например, из сходимости последовательности $(x_n(t))$ по метрике пространства $C[a,b]$ следует ее сходимость по метрике пространства $C_2[a,b]$, но обратная импликация не верна. *Пространство непрерывных функций $C[a,b]$ является полным сепарабельным пространством*. Таким же будет и метрическое пространство $L_p[a,b]$ (при $p\geq 1$) функций, измеримых на отрезке $[a,b]$ и суммируемых на нем с p-ой степенью (в смысле интеграла Лебега).

2. НОРМИРОВАННЫЕ ПРОСТРАНСТВА

Пусть наряду с множеством X элементов $x,y,\ldots$ задано какое-то числовое поле, например, множество $\mathbf{R}$ всех вещественных чисел $\alpha,\beta,\ldots$. Если в множестве X определены операции сложения (т.е. $x+y\in X$) и умножения на число (т.е. $\alpha x\in X$) так, что этим операциям присущи все свойства соответствующих операций над числами, за исключением коммутативности умножения, то говорят, что данное множество X есть ***линейное пространство*** (разновидности: линейная система, линейное множество, линеал).

Определение 8. *Линейное пространство X называется **нормированным пространством**, если каждому его элементу x ставится в соответствие вещественное число, называемое **нормой** и обозначаемое $\|x\|$, такое, что оно удовлетворяет следующим трем аксиомам (неотрицательности, однородности и треугольника):*

1) $\|x\|\geq 0$, *причем* $\|x\|=0 \Leftrightarrow x=0$;
2) $\|\lambda x\|=|\lambda|\cdot\|x\| \quad \forall\,\lambda\in\mathbf{R}$;
3) $\|x+y\|\leq\|x\|+\|y\| \quad \forall\, y\in X$.

Как и метрика, норма на заданном множестве не единственна.

Теорема 1. *Всякое нормированное пространство является метрическим.*

Действительно, пусть X — нормированное пространство с элементами $x,y,z,\ldots$ и нормой $\|\cdot\|$. Покажем, что равенство

$$\rho(x,y)=\|x-y\| \qquad (2)$$

определяет метрику ρ в пространстве X, индуцированную (ассоциированную с) данной нормой $\|\cdot\|$. Привлекая аксиомы нормы, проверяем

справедливость аксиом метрики для (2):

1) $\rho(x, y) = \|x - y\| \geq 0$, при этом

$\rho(x, y) = 0 \quad \Leftrightarrow \quad \|x - y\| = 0 \quad \Leftrightarrow \quad x - y = 0 \quad \Leftrightarrow x = y$;

2) $\rho(x, y) = \|x - y\| = \|y - x\| = \rho(y, x)$, так как

$\|x - y\| = \|(-1)(y - x)\| = |-1| \cdot \|y - x\| = \|y - x\|$;

3) $\rho(x, y) = \|x - z + z - y\| \leq \|x - z\| + \|z - y\| = \rho(x, z) + \rho(z, y)$.

В силу теоремы 1, утверждения, касающиеся метрических пространств, могут быть переформулированы в терминах норм для нормированных пространств.

Если метрическое пространство с метрикой ρ линейно, то в нем естественно ввести норму равенством

$$\|x\| = \|x - 0\| = \rho(x, 0). \tag{3}$$

Это равенство позволяет трактовать норму элемента x как расстояние от него до нуля (наличие которого обязательно в линейном пространстве).

Согласно равенству (2) и определению 2 сходимости по метрике, сходимость последовательности (x_n) к x в нормированном пространстве X означает ***сходимость по норме***, ибо

$$\|x_n - x\| \to 0 \quad \Leftrightarrow \quad \rho(x_n, x) \to 0 \quad \Leftrightarrow \quad x_n \to x.$$

Аналогично, фундаментальность последовательности (x_n) в нормированном пространстве X характеризуется требованием $\|x_n - x_m\| \xrightarrow[n, m \to \infty]{} 0$.

Определение 9. *Нормированное пространство называется* ***полным***, *если в нем всякая фундаментальная последовательность сходится. Полное нормированное пространство называется* ***банаховым*** *(или* ***B–пространством****)*[*].

Говоря о сходимости по норме, подчеркнем, что в связи с неоднозначностью введения нормы в линейном пространстве, подобно эквивалентности метрик, определяется эквивалентность норм и доказывается, что *во всяком конечномерном линейном пространстве все нормы эквивалентны.*

Примеры нормированных пространств. 1) Множество $\mathbf{R}_n$ всех n-мерных векторов с вещественными координатами, рассматривавшееся в примере 1) предыдущего пункта, с определенными в нем естественным для векторной алгебры образом операциями сложения и умножения на число образует линейное пространство, на основе которого конструируются различные нормированные пространства, например, в соответствии с введенными там метриками а), б), в). Пользуясь равенством (3), из этих

[*] В названии пространства увековечена память о польском математике Стефане Банахе (1892–1945) — одном из создателей современного функционального анализа.

метрик получаем соответственно следующие нормы для произвольного элемента $\mathbf{x} = (x_1, \ldots, x_n)$ из $\mathbf{R}_n$:

а) $\|\mathbf{x}\|_\infty = \max\limits_{i \in \{1, \ldots, n\}} |x_i|$ (*норма-максимум*);

б) $\|\mathbf{x}\|_2 = \sqrt{\sum\limits_{i=1}^{n} x_i^2}$ (*евклидова норма*);

в) $\|\mathbf{x}\|_1 = \sum\limits_{i=1}^{n} |x_i|$ (*норма-сумма*).

Иногда к этим нормам применяют геометрические названия *кубическая*, *сферическая* и *октаэдрическая* в соответствии с видом поверхности, определяемой уравнением $\|x\| = const$, когда x считается переменным радиус-вектором в трехмерном пространстве.

Получаемые с помощью норм а)–в) пространства n-мерных векторов — банаховы. Банаховым пространством $\mathbf{R}_n$ является и в более общей ситуации, когда норма вводится равенством

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} \quad (l_p\text{-норма или норма Гёльдера}) \tag{4}$$

с произвольным фиксированным $p \geq 1$. Записанные выше нормы представляют собой частные или предельный случаи l_p-нормы (а) $p \to \infty$, б) $p = 2$, в) $p = 1$).

2) Множество бесконечных последовательностей, суммируемых с p-й степенью, после введения в нем операций сложения и умножения на число по аналогии с такими же операциями для n-мерных векторов превращается в линейное пространство. Это пространство — банахово при любом $p \geq 1$ в выражении нормы

$$\|x\|_p = \left(\sum_{i=1}^{\infty} |x_i|^p \right)^{\frac{1}{p}},$$

ассоциированной с метрикой (1) и предельной по n по отношению к норме (4) в $\mathbf{R}_n$. Обозначают такое семейство нормированных пространств так же, как и соответствующее метрическое, — l_p или l^p.

3) Рассмотренные в предыдущем пункте метрические пространства непрерывных на отрезке $[a, b]$ функций легко превращаются в нормированные (для последних сохраняют те же обозначения) введением с помощью равенства (3) норм

$$\|x\| = \max_{t \in [a, b]} |x(t)|, \quad \|x\| = \sqrt{\int_a^b x^2(t)dt}, \quad \text{и} \quad \|x\| = \int_a^b |x(t)|dt.$$

Если функции $x(t)$ измеримы на $[a, b]$ и $\int_a^b |x(t)|^p dt < \infty$, где $p \geq 1$, то на

множестве таких функций строится линейное нормированное пространство $L_p[a,b]$ с нормой

$$\|x\| = \left(\int_a^b |x(t)|^p\, dt\right)^{\frac{1}{p}}$$

(в выражении которой фигурирует интеграл Лебега).

Интересно отметить тот факт, что, например, из двух нормированных пространств, $C_{L_2}[a,b]$ и $L_2[a,b]$, определяемых одной и той же по виду нормой

$$\|x\| = \left(\int_a^b x^2(t)dt\right)^{\frac{1}{2}}, \qquad (5)$$

полным, т.е. банаховым является, только одно: $L_2[a,b]$.

Важным примером сепарабельного банахова пространства является также пространство $C_k[a,b]$ (иначе, $C^k[a,b]$) k раз непрерывно дифференцируемых на $[a,b]$ функций $x(t)$ с нормой

$$\|x\| = \sum_{i=1}^{k} \max_{t\in[a,b]} \left|x^{(i)}(t)\right| \quad (\text{где } x^{(0)}(t) := x(t)),$$

частным случаем которого при $k = 0$ является пространство непрерывных функций $C[a,b]$.

3. ГИЛЬБЕРТОВЫ ПРОСТРАНСТВА

Определение 9. *Линейное пространство X называется* ***предгильбертовым*** *или* ***пространством со скалярным произведением****, если любой паре его элементов x, y можно поставить в соответствие число, в общем случае комплексное, называемое* ***скалярным*** *(или внутренним)* ***произведением*** *и обозначаемое (x,y), $\langle x,y\rangle$ или просто xy, удовлетворяющее следующим требованиям:*

1) $(x,y) = \overline{(y,x)}$;
2) $(x+y,z) = (x,z)+(y,z) \quad \forall\, z \in X$;
3) $(\lambda x, y) = \lambda(x,y) \quad \forall\, \lambda \in \mathbf{C}$;
4) $(x,x) \ge 0$, причем $(x,x) = 0 \iff x = 0$.

Предгильбертовы пространства подразделяют на ***унитарные***, что полностью соответствует определению 9, где нужно лишь уточнить, что пространство X рассматривается над полем $\mathbf{C}$ комплексных чисел, и ***евклидовы***. Последние выделяются из этого общего определения вещественностью числового поля, самого скалярного произведения (x,y) (а значит, ненужностью знака комплексного сопряжения в требовании 1)) и числа λ в требовании 3).

В пространствах со скалярным произведением $(\cdot,\cdot)$ любые два элемента x, y связаны ***неравенством Коши–Буняковского***

$$|(x,y)|^2 \le (x,x)(y,y).$$

Легко проверить, что $\sqrt{(x,x)}$ удовлетворяет всем аксиомам нормы. Таким образом, в предгильбертовом пространстве X со скалярным произведением $(\cdot,\cdot)$ естественно вводится норма

$$\|x\| := \sqrt{(x,x)}$$

(называемая *нормой, порожденной скалярным произведением*), и значит, *всякое предгильбертово пространство можно считать нормированным.* В таком случае неравенство Коши–Буняковского можно переписать в виде неравенства

$$|(x,y)| \le \|x\|\cdot\|y\|,$$

называемого также ***неравенством Коши–Шварца***.

Наличие скалярного произведения в предгильбертовых пространствах позволяет «измерять» в них не только «расстояния», но и «углы», и переносить на эти абстрактные пространства многие геометрические свойства реальных трехмерных пространств. Например, известное соотношение между длинами сторон и диагоналей параллелограмма обобщается на предгильбертовы пространства в виде так называемого ***равенства параллелограмма***

$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2.$$

Сходимость в пространствах со скалярными произведениями определяется, в основном, как сходимость по норме, т.е.

$$x_n \to x, \quad \text{если } \|x_n - x\| = \sqrt{(x_n - x, x_n - x)} \to 0.$$

Ее называют ***сильной сходимостью*** (есть еще ***слабая сходимость***:

$$x_n \to x, \quad \text{если } (x_n, y) \to (x,y) \quad \forall\, y \in X).$$

Определение 10. *Предгильбертово пространство называется* ***гильбертовым****, если оно полно в норме, порожденной скалярным произведением (т.е. в смысле сильной сходимости).*

Гильбертовы пространства (названные так в честь знаменитого ученого Гильберта*) и обычно обозначаемые буквой H), являясь одновременно банаховыми, играют особо важную роль в функциональном анализе и в вычислительной математике. Одной из предпосылок для этого служит возможность трактовать величину $\dfrac{(x,y)}{\|x\|\cdot\|y\|} \in [-1,1]$ как косинус угла между элементами x и y и, благодаря этому, проектировать одни элементы на другие.

Определение 11. *Два элемента x и y из гильбертова пространства H называются* ***ортогональными*** *(записывают $x \perp y$), если $(x,y) = 0$. Элемент $x \in H$ ортогонален множеству $M \subset H$ (записывают $x \perp M$), если x ортогонален каждому элементу y из M.*

*) Гильберт Давид (Hilbert David, 1862–1943) — крупнейший немецкий ученый, оказавший большое влияние на развитие многих разделов математики.

Теорема 2. *Пусть x — элемент гильбертова пространства H, а M — некоторое подпространство H (линейное замкнутое множество элементов из H). Тогда найдется единственная пара элементов $y \in M$ и $z \perp M$ таких, что $x = y + z$.*

Элемент y в этой ***теореме о разложении*** расценивается как ортогональная ***проекция*** произвольно заданного элемента x гильбертова пространства H на его подпространство M, а элемент z из ортогонального дополнения $M^{\perp}$ подпространства M до H — как «расстояние» от x до M. Так как $M^{\perp}$, в свою очередь, является подпространством H, то y и z и, соответственно, M и $M^{\perp}$ можно поменять ролями.

Следствием теоремы 2 в ее условиях является равенство $\|x\|^2 = \|y\|^2 + \|z\|^2$, называемое ***теоремой Пифагора***.

Наиболее типичными примерами гильбертовых пространств могут служить следующие пространства*).

1) Евклидово пространство — пространство n-мерных векторов $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{y} = (y_1, \ldots, y_n)$, ... с вещественными координатами, где скалярное произведение $(\mathbf{x}, \mathbf{y})$ определяется, как и в реальных пространствах, равенством $(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} x_i y_i$ и соответствует евклидовой норме. Это пространство часто обозначают E_n.

2) Пространство l_2 бесконечных последовательностей $\mathbf{x} = (x_1, \ldots, x_n, \ldots)$, $\mathbf{y} = (y_1, \ldots, y_n, \ldots)$, ... с вещественными или комплексными членами, для которых имеет место сходимость рядов из квадратов модулей членов. Скалярное произведение здесь вводится равенством $(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} x_i y_i$ или равенством $(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} x_i \overline{y}_i$ соответственно в вещественном или комплексном случаях.

3) Пространство $L_2[a, b]$ вещественнозначных функций $x(t)$, $y(t)$, ..., определенных при $t \in [a, b]$, измеримых и суммируемых с квадратом по Лебегу. Скалярное произведение в $L_2[a, b]$ определяется равенством $(x, y) = \int_a^b x(t) y(t) dt$ и порождает норму (5).

4. ЛИНЕЙНЫЕ ОПЕРАТОРЫ

Под ***оператором*** A понимается отображение, т.е. правило, по которому элементу x одного множества X ставится в соответствие элемент $y = Ax$ того же или другого множества Y. Употребительна запись $A : X \to Y$.

---

*) Первый пример противоречит даваемым иногда определениям гильбертовых пространств как заведомо бесконечномерных.

Пусть X и Y — два линейных пространства над одним и тем же числовым полем K и A — оператор, ставящий в соответствие элементам x линейного подпространства $D(A)$ пространства X элементы $y \in Y$.

Определение 12. *Оператор A называется* ***аддитивным****, если*

$$A(x_1 + x_2) = Ax_1 + Ax_2 \quad \forall x_1,\ x_2 \in D(A).$$

Оператор A называется ***однородным****, если*

$$A(\lambda x) = \lambda Ax \quad \forall x \in D(A) \quad \text{и} \quad \forall \lambda \in K.$$

Оператор A называется ***линейным****, если он аддитивен и однороден (иначе, если он* ***дистрибутивен****, т.е. если $A(\lambda_1 x_1 + \lambda_2 x_2) = \lambda_1 Ax_1 + \lambda_2 Ax_2$.*

Будем далее считать X и Y нормированными пространствами.

Определение 13. *Аддитивный оператор $A : X \to Y$ называется* ***ограниченным****, если существует такая постоянная C, что при любых $x \in D(A) \subseteq X$ выполняется неравенство*

$$\|Ax\| \le C\|x\|.$$

Наименьшая из таких постоянных C называется ***нормой*** *оператора A и обозначается $\|A\|$.*

Таким образом, имеет место неравенство

$$\|Ax\| \le \|A\| \cdot \|x\|, \tag{6}$$

которое называют ***условием согласованности норм***.

Определение 14. *Оператор $A : X \to Y$ называется* ***непрерывным в точке*** *$x \in D(A)$, если для любой последовательности точек $x_n \in D(A)$ таких, что $x_n \to x$, справедливо $Ax_n \to Ax$, т.е.*

$$\|x_n - x\| \to 0 \quad \Rightarrow \quad \|Ax_n - Ax\| \to 0.$$

Характерно, что для линейных операторов непрерывность в одной точке области определения влечет непрерывность в любой другой точке этой области, и, кроме того, *непрерывность линейных операторов равнозначна их ограниченности*. В связи с этим, часто свойство непрерывности или ограниченности оператора включают в определение его линейности.

Норма оператора $A : X \to Y$, удовлетворяющая условию согласованности норм (6), может быть получена с помощью норм пространств X и Y на основе следующих равенств:

$$\|A\| = \sup_{x \ne 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\| \le 1} \|Ax\| = \sup_{\|x\| = 1} \|Ax\| \tag{7}$$

(называемых иногда ***условиями подчиненности норм***).

Введя над операторами линейные операции (суммой определенных на X операторов A и B называется такой оператор $U := A + B$, что $Ux = Ax + Bx \ \forall x \in X$, а произведением оператора A на число λ — такой оператор $V := \lambda A$, что $Vx = \lambda Ax \quad \forall x \in X$), можно показать, что *множест-*

во всех линейных непрерывных операторов из X в Y само образует линейное нормированное пространство, которое обычно обозначают $L(X,Y)$ или $[X \to Y]$, причем нормой элементов этого пространства может служить операторная норма (7). Если пространство Y — полное, то и пространство $L(X,Y)$ — полное, т.е. банахово.

Определив произведение операторов $A: Y \to Z$ и $B: X \to Y$ как оператор $C := AB: X \to Z$ такой, что $Cx = ABx = A(Bx) \quad \forall x \in X$, можно убедиться, что если A и B — линейные непрерывные, то и оператор AB будет таким же, причем

$$\|AB\| \le \|A\| \cdot \|B\|, \tag{8}$$

Пусть X — нормированное пространство и I — тождественный оператор в X (т.е. $Ix = x \quad \forall x \in X$). Тогда, если A и B принадлежат пространству $L(X,X)$, то и $AB \in L(X,X)$, и это пространство линейных непрерывных операторов образует нормированное кольцо. В нем естественным образом определяются натуральные степени оператора и, как следствие ***условия мультипликативности операторных*** норм (8), имеет место неравенство

$$\|A^n\| \le \|A\|^n \quad \forall n \in \mathbf{N}_0$$

(по определению считается $A^0 := I$, а из (7) следует $\|I\| = 1$). Наличие в нормированных кольцах степенных операторов позволяет рассматривать в них степенные ряды и, в частности, «геометрическую прогрессию»

$$I + A + A^2 + \ldots + A^k + \ldots. \tag{9}$$

Определение 15. *Пусть A — линейный оператор из X в Y, а I_X и I_Y — тождественные операторы в пространствах X и Y соответственно. Операторы V и U, действующие из Y в X, такие, что $VA = I_X$, $AU = I_Y$, называются соответственно* ***левым*** *и* ***правым обратными операторами***. *Если левый и правый обратные операторы одновременно существуют, то они совпадают и называются двусторонними обратными или просто* ***обратным*** *для A оператором A^{-1}.*

Существование правого обратного для A оператора означает существование решения операторного уравнения $Ax = y$ без гарантии единственности, а существование левого — наоборот, обеспечивает единственность, но не дает гарантии существования решения. Ясно, что если оператор A имеет обратный A^{-1}, то решение x этого уравнения существует, единственно и представимо в виде $x = A^{-1}y$.

Наличие ограниченного обратного оператора A^{-1} для $A \in L(X,Y)$ связывают с существованием такой постоянной $\delta > 0$, что

$$\|Ax\| \ge \delta \|x\| \quad \forall\, x \in X;$$

при этом $\|A^{-1}\| \le 1/\delta$.

Один из достаточных признаков обратимости оператора $I - A$ имеет следующую формулировку.

Теорема 3 (Банаха). *Пусть X — банахово пространство и $A \in L(X,X)$. Тогда, если $\|A\| \le q < 1$, то оператор $I - A$ имеет непрерывный обратный оператор $(I-A)^{-1}$, являющийся суммой ряда* (9), *причем*

$$\left\|(I-A)^{-1}\right\| \le \frac{1}{1-q}.$$

Наиболее простой и важный для вычислительной математики пример линейного оператора — это линейное преобразование векторов-столбцов $\mathbf{x}$ пространства $\mathbf{R}_n$ в векторы-столбцы $\mathbf{y}$ пространства $\mathbf{R}_m$. Каждое такое линейное преобразование $\mathbf{y} = \mathbf{Ax}$ однозначно определяется $m \times n$-матрицей $\mathbf{A}$ вида

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix},$$

которую можно отождествить с линейным оператором $A: R_n \to R_m$.

Теорема 4. *Пусть A — линейный оператор из $\mathbf{R}_n$ в $\mathbf{R}_m$ и в пространствах $\mathbf{R}_n$ и $\mathbf{R}_m$ введены одинаковые l_p-нормы, где $p = 1$, $p = 2$ и $p = \infty$ (норма-сумма, евклидова норма и норма-максимум). Тогда $A \in L(R_n, R_m)$, и соответствующие (согласованные и подчиненные) нормы матрицы $\mathbf{A}$ определяются равенствами:*

при $p = 1$ $\quad \|\mathbf{A}\|_1 = \max\limits_{j \in \{1,\ldots,n\}} \sum\limits_{i=1}^{m} |a_{ij}|$;

при $p = 2$ $\quad \|\mathbf{A}\|_2 = \sqrt{\Lambda}$, *где Λ — наибольшее собственное число матрицы* $\mathbf{A}^T\mathbf{A}$;

при $p = \infty$ $\quad \|\mathbf{A}\|_\infty = \max\limits_{i \in \{1,\ldots,m\}} \sum\limits_{j=1}^{n} |a_{ij}|$.

Иногда наряду со ***спектральной нормой*** $\|\mathbf{A}\|_2$ применяют другую матричную норму, также согласованную с евклидовой нормой вектора, но не подчиненную ей, и, как следствие, немультипликативную, — ***норму Фробениуса***

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}$$

(другие ее названия: ***евклидова***, ***шуровская***, ***Э. Шмидта***).

Пусть A — линейное преобразование в пространстве $\mathbf{R}_n$ и $\mathbf{A}$ — квадратная матрица этого преобразования. Справедливо неравенство $|\lambda| \le \|\mathbf{A}\|$, где λ — произвольное собственное число матрицы $\mathbf{A}$, а $\|\cdot\|$ — произвольная матричная норма.

5. ФУНКЦИОНАЛЫ И СОПРЯЖЕННЫЕ ОПЕРАТОРЫ

Функционал — это оператор, значениями которого являются вещественные или комплексные числа. Примерами функционалов могут служить определенный интеграл, метрика, норма, скалярное произведение.

Ограничимся рассмотрением вещественнозначных линейных функционалов. Так как они являются частными случаями линейных операторов, то к ним автоматически могут быть отнесены все результаты, известные для линейных операторов.

Если функционал $f: X \to \mathbf{R}$, где X — вещественное нормированное пространство, ограниченный, то в соответствии с (7) его норма есть

$$\|f\| = \sup_{\|x\| \le 1} |f(x)|.$$

Одним из центральных результатов теории линейных функционалов является следующая ***теорема о продолжении***.

Теорема 5 (**Хана–Банаха**). *Всякий линейный ограниченный функционал f, заданный на подпространстве D нормированного пространства X, можно продолжить на все пространство X с сохранением нормы (т.е. существует определенный на всем пространстве X такой линейный ограниченный функционал $\tilde{f}$, что $\|\tilde{f}\| = \|f\|$ и $\tilde{f}(x) = f(x)$ $\forall x \in D$).*

Другой важный результат касается представления линейных функционалов и их норм в гильбертовых пространствах.

Теорема 6 (**Рисса**). *Пусть в гильбертовом пространстве H задан линейный ограниченный функционал f. Тогда найдется элемент $y \in H$, однозначно определяемый функционалом f, такой, что $f(x) = (x, y)$ и $\|f\| = \|y\|$.*

Так, в евклидовом пространстве $\mathbf{R}_n$ n-мерных векторов $\mathbf{x} = (x_1, \ldots, x_n)$ с вещественными координатами x_i любой линейный ограниченный функционал имеет вид $f(\mathbf{x}) = (\mathbf{x}, \mathbf{c}) = \sum_{i=1}^{n} c_i x_i$ с нормой $\|f\| = \sqrt{\sum_{i=1}^{n} c_i^2}$, где $\mathbf{c} = (c_1, \ldots, c_n)$ — фиксированный вектор из $\mathbf{R}_n$ (свой для каждого функционала f). Аналогично, в пространстве $L_2[a, b]$ измеримых и суммируемых с квадратом на $[a, b]$ действительных функций $x(t)$ линейный функционал и его норма, соответственно,

$$f(x) = (x, y) = \int_a^b x(t) y(t) dt \quad \text{и} \quad \|f\| = \sqrt{\int_a^b y^2(t) dt},$$

где $y(t) \in L_2[a, b]$ — фиксированная функция.

Множество линейных ограниченных функционалов, определенных на нормированном пространстве X, является банаховым пространством (независимо от полноты X), которое называют ***сопряженным к X пространством*** и обозначают X^*. Если пространство X^* совпадает с X, то пространство X называют ***самосопряженным***.

С помощью понятий линейного функционала и сопряженного пространства в нормированных пространствах вводится понятие сопряженного к линейному оператору $A: X \to Y$ оператора $A^*: Y^* \to X^*$ и доказывается, что $\|A^*\| = \|A\|$.

Более просто сопряженный к A оператор A^* определяется в случае, когда A действует из H в H, где H — гильбертово пространство.

Определение 16. *Линейный оператор A^* называется **сопряженным** к линейному оператору $A: H \to H$, если*

$$(Ax, y) = (x, A^* y) \quad \forall x, y \in H.$$

Определение 17. *Линейный ограниченный оператор A называется **самосопряженным** (или **эрмитовым**), если $A^* = A$, т.е. если*

$$(Ax, y) = (x, Ay) \quad \forall\, x, y \in H.$$

Заметим, что $(A^*)^* = A$, $(A^*)^{-1} = (A^{-1})^*$, а операторы A^*A и AA^* — самосопряженные ***неотрицательные операторы*** (неотрицательность их понимается в смысле $(A^*Ax, x) \ge 0$ и $(AA^*x, x) \ge 0$ $\forall\, x \in H$).

Для самосопряженного оператора A имеет место равенство $\|A^n\| = \|A\|^n$, в силу чего *условие* $\|A\| < 1$ *является не только достаточным, но и необходимым для сходимости операторной геометрической прогрессии* (9). При этом норма самосопряженного оператора A может быть представлена через его ***границы***

$$m_A := \inf_{\|x\|=1} (Ax, x) \quad \text{и} \quad M_A := \sup_{\|x\|=1} (Ax, x)$$

следующим образом:

$$\|A\| = \sup_{\|x\|=1} |(Ax, x)| = \max\{|m_A|, |M_A|\}.$$

Как и для матриц, для самосопряженных операторов определяются собственные числа и собственные элементы; показывается, что все собственные числа самосопряженного оператора A вещественны и находятся между его границами m_A и M_A, а соответствующие различным собственным числам собственные элементы взаимно ортогональны.

С неотрицательным самосопряженным оператором A в вещественном гильбертовом пространстве H справедливо ***обобщенное неравенство Коши–Буняковского***

$$(Ax, y)^2 \le (Ax, x)(Ay, y) \quad \forall\, x, y \in H ,$$

а в случае положительности самосопряженного оператора A (в смысле $(Ax, x) > 0 \quad \forall x \ne 0$) и его ограниченности имеет место оценка

$$\|Ay\|^2 \le \|A\| \cdot (Ay, y) \quad \forall\, y \in H .$$

В конечномерном евклидовом пространстве $\mathbf{R}_n$, где линейный оператор $A : \mathbf{R}_n \to \mathbf{R}_n$ отождествляется с квадратной матрицей $\mathbf{A}$ размерности n, роль сопряженного к A оператора A^* выполняет матрица $\mathbf{A}^{\mathrm{T}}$, транспонированная по отношению к $\mathbf{A}$. Следовательно, самосопряженными линейными операторами в $\mathbf{R}_n$ следует считать линейные преобразования, осуществляемые посредством симметрических матриц, ибо для них $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$. Свойства таких матриц хорошо известны.

Если H — произвольное гильбертово пространство и M — некоторое его подпространство, то оператор P, ставящий в соответствие каждому элементу $x \in H$ единственный элемент $y \in M$ (его существование обусловлено теоремой 2), называют ***оператором ортогонального проектирования*** или ***ортопроектором***. Этот оператор P обладает свойствами $P \in L(H, H)$, $\|P\| = 1$, $P^2 = P$ и является самосопряженным оператором.

6. НЕЛИНЕЙНЫЕ ОПЕРАТОРЫ В НОРМИРОВАННЫХ ПРОСТРАНСТВАХ

В классическом математическом анализе и его приложениях к вычислительной математике многого удается достичь благодаря линеаризации — замене нелинейных функций близкими линейными. В определенных случаях такая локальная замена осуществляется с помощью производной. Техника линеаризации успешно переносится на нелинейные операторы в нормированных пространствах.

Пусть $F(x)$ — нелинейный оператор, действующий из банахова пространства X в банахово пространство Y, и пусть точка x_0 принадлежит его области определения $D(F)$ вместе с некоторой окрестностью S.

Определение 17. *Оператор $F(x)$ называется **дифференцируемым по Фреше** в точке x_0, если существует такой линейный непрерывный оператор $A \in L(X, Y)$, что при любых $h \in X$, при которых $x_0 + h \in S$,*

$$F(x_0 + h) - F(x_0) = Ah + \omega(x_0, h), \tag{10}$$

где $\|\omega(x_0, h)\| = o(\|h\|)$ при $h \to 0$. Оператор A в представлении (10) *называется **производной (Фреше)** и обозначается $F'(x_0)$, а выражение $Ah = F'(x_0)h = dF(x_0, h)$ — **дифференциалом (Фреше)**.*

Производную и дифференциал Фреше называют еще соответственно ***сильной производной*** и ***сильным дифференциалом*** в отличие от также существующих слабых поизводной и дифференциала (Гато).

Если посмотреть на векторную функцию $\mathbf{F}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$ векторного аргумента $\mathbf{x} = (x_1, \ldots, x_n)$ как на нелинейный оператор, осуществляющий отображение точек n-мерного вещественного пространства $\mathbf{R}_n$ в m-мерное вещественное пространство $\mathbf{R}_m$, нетрудно убедиться, что производной Фреше такого оператора служит матрица частных производных (***матрица Якоби***)

$$\mathbf{F}'(\mathbf{x}) = \mathbf{J}(\mathbf{x}) := \begin{pmatrix} \dfrac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \dfrac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \cdot\ \cdot\ \cdot & \cdot\ \cdot\ \cdot & \cdot\ \cdot\ \cdot \\ \dfrac{\partial f_m(\mathbf{x})}{\partial x_1} & \cdots & \dfrac{\partial f_m(\mathbf{x})}{\partial x_n} \end{pmatrix}.$$

В частности, если $F : \mathbf{R}_n \to \mathbf{R}_1$, т.е. если $F(\mathbf{x}) = f(x_1, \ldots, x_n)$ — функция n переменных, то

$$\mathbf{F}'(\mathbf{x}) = \left(\frac{\partial f_1(\mathbf{x})}{\partial x_1}, \quad \ldots, \quad \frac{\partial f_1(\mathbf{x})}{\partial x_n} \right) = \operatorname{grad} F .$$

Таким образом, ***градиент*** — это матрица-строка Якоби.

Производная Фреше нелинейного оператора обладает рядом свойств, присущих производным функций одной или нескольких переменных. Отметим некоторые из них.

1) Если оператор $F(x)$ дифференцируем в точке x_0, то он непрерывен в ней.

2) Производная постоянного оператора (например, λI где $\lambda \in \mathbf{R}$) равна нулю.

3) Если $F(x) \equiv Ax$, где $A \in L(X, Y)$, то $F'(x) \equiv A$.

4) Если $F_1(x)$ и $F_2(x)$ — дифференцируемые в точке x_0 операторы, то при любых $\alpha_1, \alpha_2 \in \mathbf{R}$ оператор $F(x) := \alpha_1 F_1(x) + \alpha_2 F_2(x)$ дифференцируем в этой точке и $F'(x_0) := \alpha_1 F'(x_0) + \alpha_2 F'(x_0)$.

5) Если оператор $P : X \to Y$ дифференцируем в точке $x_0 \in X$, а оператор $Q : Y \to Z$ дифференцируем в точке $y_0 = P(x_0) \in Y$, то оператор $F := QP : X \to Z$ дифференцируем в точке x_0, причем $F'(x_0) = Q'(y_0)P'(x_0)$.

Пусть $F(t)$, где $t \in [a, b]$ — непрерывная абстрактная функция со значениями в банаховом пространстве Y. Для нее существует ***абстрактный интеграл***

$$\int_a^b F(t)dt := \lim_{\substack{n \to \infty \\ \max \Delta t_k \to 0}} \sum_{k=1}^{n} F(\tau_k)\Delta t_k ,$$

где $\Delta t_k := t_k - t_{k-1}$, t_k — точки разбиения отрезка $[a, b]$ на n частей,

а τ_k — произвольные точки промежутков $[t_{k-1}, t_k]$. Из свойств этого интеграла выделим три.

1) Если A — линейный непрерывный оператор, то

$$\int_a^b AF(t)dt = A\int_a^b F(t)dt.$$

2) Если $F(t) = \varphi(t)y_0$, где $\varphi(t)$ — непрерывная на $[a, b]$ вещественная функция, а y_0 — фиксированный элемент банахова пространства Y, то

$$\int_a^b F(t)dt = y_0\int_a^b \varphi(t)dt.$$

3) $\left\|\int_a^b F(t)dt\right\| \le \int_a^b \|F(t)\|\, dt$.

Через абстрактный интеграл вводится ***интеграл по отрезку***:

$$\int_{x_0}^{x_0+\Delta x} R(x)dx := \int_0^1 R(x_0 + t\Delta x)dt\, \Delta x; \qquad (11)$$

здесь $R(x)$ — оператор, определенный на выпуклом множестве M банахова пространства X, содержащем *отрезок* $[x_0, x_0 + \Delta x]$, со значениями в банаховом пространстве $L(X, Y)$.

Для непрерывно дифференцируемого оператора имеет место ***формула Ньютона–Лейбница***

$$\int_{x_0}^{x_0+\Delta x} F'(x)dx = F(x_0 + \Delta x) - F(x_0).$$

С учетом (11) последняя может быть переписана в виде равенства

$$F(x_0 + \Delta x) - F(x_0) = \int_0^1 F'(x_0 + \Theta\Delta x)d\Theta\, \Delta x,$$

которое называют ***формулой конечных приращений Лагранжа*** в интегральной форме.

Формулой конечных приращений называют также неравенство

$$\|F(x_0 + \Delta x) - F(x_0)\| \le \sup_{\Theta\in(0,1)} \|F'(x_0 + \Theta\Delta x)\| \cdot \|\Delta x\|,$$

а неравенство

$$\|F(x_0 + \Delta x) - F(x_0) - F'(x_0)\Delta x\| \le \|\Delta x\| \cdot \sup_{\Theta\in(0,1)} \|F'(x_0 + \Theta\Delta x) - F'(x_0)\|$$

носит название ***формула конечных приращений с остаточным членом***.

Предположим, что нелинейный оператор $F(x): X \to Y$ дифференцируем во всех точках $x \in X$. Тогда при каждом фиксированном x оператор $F'(x): X \to L(X, Y)$ линеен. Если же x считать переменным элементом X, то $F'(x)$, вообще говоря, — нелинейный оператор, и для него можно вновь ставить вопрос о дифференцировании. Таким образом, приходим к понятию ***второй производной оператора*** $F(x)$, каковой считаем оператор

$$F''(x) := (F'(x))' : \quad X \to L(X, L(X, Y)).$$

Вторая производная является представителем билинейных операторов.

Определение 18. *Оператор B, ставящий паре элементов x, $\tilde{x} \in X$ элемент $y = B(x, \tilde{x}) \in Y$, называется **билинейным**, если он линеен по каждому из аргументов, т.е.*

$$B(\alpha_1 x_1 + \alpha_2 x_2, \tilde{x}) = \alpha_1 B(x_1, \tilde{x}) + \alpha_2 B(x_2, \tilde{x}),$$
$$B(x, \beta_1\tilde{x}_1 + \beta_2\tilde{x}_2) = \beta_1 B(x, \tilde{x}_1) + \beta_2 B(x, \tilde{x}_2),$$

и ограничен в совокупности, т.е.

$$\|B(x, \tilde{x})\| \le C\|x\| \cdot \|\tilde{x}\| \quad \forall\, x, \tilde{x} \in X$$

(*наименьшая из таких постоянных $C > 0$ называется **нормой** билинейного оператора B*).

Множество билинейных операторов, действующих из нормированного пространства X в банахово пространство Y, образует банахово пространство $L(X^2, Y)$, изометричное пространству $L(X, L(X, Y))$. Следовательно, $F''(x) \in L(X^2, Y)$, причем $F''(x)hh = F''(x)h^2$.

Обращаясь к примеру m-мерной векторной функции $\mathbf{F}(\mathbf{x})$ векторного n-мерного аргумента $\mathbf{x}$, нетрудно выяснить, что ее вторая производная есть отображение пространства $\mathbf{R}_n \times \mathbf{R}_n$ в $\mathbf{R}_m$, и результат такого отображения — m-мерный вектор $\mathbf{F}''(\mathbf{x})\mathbf{hg}$ при произвольных $\mathbf{h}$ и $\mathbf{g}$ из $\mathbf{R}_n$ имеет представление

$$\mathbf{F}''(\mathbf{x})\mathbf{hg} = \left(\mathbf{g}^{\mathrm{T}}\mathbf{H}_1(\mathbf{x})\mathbf{h}, \ldots, \mathbf{g}^{\mathrm{T}}\mathbf{H}_m(\mathbf{x})\mathbf{h}\right),$$

где $\mathbf{H}_j(\mathbf{x})$ — ***матрица Гессе***, вид которой

$$\mathbf{H}_j(\mathbf{x}) = \begin{pmatrix} \dfrac{\partial^2 f_j(\mathbf{x})}{\partial x_1^2} & \dfrac{\partial^2 f_j(\mathbf{x})}{\partial x_1\partial x_2} & \cdots & \dfrac{\partial^2 f_j(\mathbf{x})}{\partial x_1\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{\partial^2 f_j(\mathbf{x})}{\partial x_n\partial x_1} & \dfrac{\partial^2 f_j(\mathbf{x})}{\partial x_n\partial x_2} & \cdots & \dfrac{\partial^2 f_j(\mathbf{x})}{\partial x_n^2} \end{pmatrix}.$$

Подобно билинейным операторам и вторым производным вводятся полилинейные операторы и производные k-го порядка $F^{(k)}(x) \in L(X^k, Y)$ при $k \in \mathbf{N}$. Наличие у нелинейного оператора $F(x)$ производных до $(k+1)$–го порядка позволяет записать для него ***формулу Тейлора***

$$F(x) \equiv F(x_0 + h) = F(x_0) + F'(x_0)h + \frac{1}{2}F''(x_0)h^2 + \ldots$$
$$\ldots + \frac{1}{k!}F^{(k)}(x_0)h^k + \frac{1}{k!}\int_{x_0}^{x_0+h} F^{(k+1)}(z)(x - z)^k\, dz,$$

где последнее слагаемое — одна из форм представления остаточного члена (через интеграл по отрезку).

ОБРАЗЦЫ ПОСТАНОВОК ЛАБОРАТОРНЫХ ЗАДАНИЙ

Лабораторная работа 1. **«Метод Гаусса и LU-разложение матриц»**

Дана система $\mathbf{Ax} = \mathbf{b}$, где:

вариант 1

$$\mathbf{A} = \begin{pmatrix} 14 & -8 & -21 & 12 \\ 10 & -6 & -15 & 9 \\ 35 & -20 & -56 & 32 \\ 25 & -15 & -40 & 24 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 19 \\ 14 \\ 53 \\ 39 \end{pmatrix};$$

вариант 2

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & -3 & -9 \\ 0 & 1 & -7 & -21 \\ 3 & 12 & -92 & -279 \\ 1 & 4 & -31 & -94 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 11 \\ 22 \\ 297 \\ 100 \end{pmatrix};$$

и т.д.

1. Решить систему методом Гаусса. Предусмотреть постолбцовый выбор главного элемента и итерационное уточнение решения до достижения точности $\varepsilon = 10^{-12}$ по евклидовой норме невязки в рамках применяемой схемы реализации метода*).

2. Выполнить LU-разложение матрицы $\mathbf{A}$ и с его помощью получить $\det \mathbf{A}$ и решение $\mathbf{x}$ данной системы.

3. Найти матрицу $\mathbf{X} = \mathbf{A}^{-1}$ двумя способами:
   а) решая подсистемы $\mathbf{Ax}^j = \mathbf{e}^j$ системы $\mathbf{AX} = \mathbf{E}$ (используя при этом фрагменты выполнения п.1);
   б) применяя готовые формулы, полученные на основе LU -разложения.

4. Вычислить $\operatorname{cond} \mathbf{A}$ в различных простых нормах и охарактеризовать чувствительность данной системы к погрешностям исходных данных.

---

*) Лучше взять $\varepsilon = 10 \cdot masheps$, предварительно найдя *masheps* используемого компьютера.

Лабораторная работа 2. **«Метод прогонки»**

Методом прогонки найти вектор $(u_1;\, u_2;\, \ldots;\, u_{10})$, являющийся решением уравнения (системы)

$$a_k u_{k-1} + b_k u_k + c_k u_{k+1} = f_k,$$

где $k = 1, 2, \ldots, 10$; $a_1 = c_{10} = 0$; коэффициенты a_k (при $k = 2, 3, \ldots, 10$), b_k (при $k = 1, 2, \ldots, 10$), c_k (при $k = 1, 2, \ldots, 9$) и f_k (при $k = 1, 2, \ldots, 10$) задаются следующей таблицей:

	a_k	b_k	c_k	f_k
вариант 1	k	$3.1k$	$-2k$	$\dfrac{2.1k^2 + 7.2k + 2}{k^2 + 3k + 2}$
вариант 2	$\dfrac{3}{k}$	$\dfrac{11}{10k}$	$\dfrac{2}{k}$	$30.5 - \dfrac{41.6}{k}$
…	…	…	…	…

Что можно сказать об устойчивости прогонки в данном конкретном случае?

Лабораторная работа 3. **«Прямое и итерационное решение симметричной линейной алгебраической системы»**

С точностью $\varepsilon = 10^{-12}$ найти решение системы

$$\sum_{j=1}^{6} a_{ij} x_j = b_j \quad (i = 1, 2, \ldots, 6)$$

с матрицей коэффициентов вида

$$\mathbf{A} = (a_{ij}) = \begin{pmatrix} p_1 & 0.1p_1 & 0 & 0 & q & 0 \\ 0.1p_1 & p_2 & 0.1p_2 & 0 & 0 & q \\ 0 & 0.1p_2 & p_3 & 0.1p_3 & 0 & 0 \\ 0 & 0 & 0.1p_3 & p_4 & 0.1p_4 & 0 \\ q & 0 & 0 & 0.1p_4 & p_5 & 0.1p_5 \\ 0 & q & 0 & 0 & 0.1p_5 & p_6 \end{pmatrix},$$

если:

	$p_i\ (i = 1, 2, \ldots, 6)$	$b_i\ (i = 1, 2, \ldots, 6)$	q
вариант 1	i	1	–0.5
вариант 2	$10 - i$	$25 - 9i$	2
…	…	…	…

Применить:

а) метод квадратных корней;

б) метод Якоби (сделав предварительно подсчет числа итераций, гарантирующего получение решения с заданной точностью);

в) метод Зейделя;

г) метод сопряженных градиентов.

Попытаться уменьшить число итераций метода Зейделя, вводя релаксационный параметр и оптимизируя его значение экспериментальным путем.

Провести сравнительный анализ примененных методов.

Лабораторная работа 4. **«Численное решение алгебраических проблем собственных значений»**

Дано матрица:

вариант 1 $\begin{pmatrix} 7 & -1 & -2 & 3 \\ -1 & 6 & 0 & 2 \\ -2 & 0 & 5 & 1 \\ 3 & 2 & 1 & 7 \end{pmatrix}$; *вариант* 2 $\begin{pmatrix} 5 & 2 & 0 & -1 \\ 2 & 7 & -3 & 1 \\ 0 & -3 & 9 & 4 \\ -1 & 1 & 4 & 8 \end{pmatrix}$;

и т.д.

1. Найти наибольшее по модулю собственное число и соответствующий ему собственный вектор

а) степенным методом;

б) методом скалярных произведений.

(В качестве начального взять вектор $(1;\ 1;\ 1;\ 1)$).

2. Найти наименьшее по модулю собственное число и соответствующий ему собственный вектор

а) методом обратных итераций;

б) методом обратных итераций с отношениями Рэлея.

3. Решить полную проблему собственных значений методом вращений Якоби.

Точность $\varepsilon = 10^{-6}$ (в евклидовой норме).

Лабораторная работа 5. **«Методы решения скалярных уравнений»**

С точностью $\varepsilon = 10^{-12}$ найти каждый из корней уравнения:

вариант 1 $4x\ln^2 x - 4\sqrt{1+x} + 5 = 0$;

вариант 2 $x^4 e^x + \sqrt[3]{x-1} - 2 = 0$;

и т.д.

каждым из следующих способов:

1) методом половинного деления;

2) методом хорд;

3) методом Ньютона;

4) методом секущих;

5) полюсным методом Ньбтона (с подвижным полюсом);

6) полюсным методом секущих.

Сравнить методы по числу итераций и по вычислительным затратам. Что можно сказать об эффективности примененных методов?

Лабораторная работа 6. **«Скалярная задача о неподвижной точке»**

С точностью $\varepsilon = 10^{-12}$ решить уравнение:

вариант 1 $e^{-0.45x} - \sqrt{x-3} = 0$;

вариант 2 $(x-4)^3 + \ln x = 0$;

и т.д.

а) методом простых итераций;

б) Δ^2-процессом Эйткена;

в) методом Вегстейна.

Предварительно привести уравнение к виду, пригодному для проведения итераций; доказать существование и единственность корня; выбрав начальное приближение, сделать априорную оценку количества шагов метода простых итераций.

Результаты представить по следующей форме:

Метод	Начальное приближение	Априорное число итераций	Фактическое число итераций	Полученный корень	Невязка
МПИ					
Эйткена					
Вегстейна					

Сравнить методы по требуемому количеству вычислений функций для получения решения с заданной точностью.

Лабораторная работа 7. **«Метод Бернулли вычисления корней многочлена»**

Дано уравнение:

вариант 1 $x^4 + 1.1x^3 - 11.51x^2 - 2.331x - 0.117 = 0$;

вариант 2 $x^4 - 10.3x^3 - 15.75x^2 - 286.875x - 84.375 = 0$;

и т.д.

1. Непосредственным применением метода Бернулли найти наибольший и наименьший по модулю корни.

2. Используя найденные корни, понизить по схеме Горнера степень многочлена и найти остальные корни.

Лабораторная работа 8. **«Решение систем нелинейных уравнений»**

Дана система и начальная точка:

вариант 1 $\begin{cases}(x-y)^3-8(x+y)=0, & x_0=2,\\ 2(x-y)+15\ln(x+y)-5=0, & y_0=-0.5;\end{cases}$

вариант 2 $\begin{cases}0.8x^2+2xy+1.3y^2+20x-15y=0, & x_0=0.5,\\ e^{0.6y-0.8x}-1.14x-1.52y=0, & y_0=1;\end{cases}$

и т.д.

Найти решение данной системы, исходя из данной начальной точки, следующими методами:

1) основным методом Ньютона (явным или неявным);
2) разностным методом Ньютона (с разными шагами дискретизации производной);
3) модифицированным (упрощенным) методом Ньютона;
4) методом Ньютона с аппроксимацией обратных матриц;
5) методом Брауна;
6) методом секущих Бройдена;
7) методом градиентного спуска.

Провести сравнение всех указанных методов решения нелинейных систем на основе конкретного вычислительного материала, полученного при задании точности $\varepsilon=10^{-3}$, 10^{-6}, 10^{-12}.

Лабораторная работа 9. **«Интерполяция»**

Многократно дифференцируемая функция $y=f(x)$ задана таблицей значений

	x	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6
вариант 1	y	1	1.0201	1.0811	1.1855	1.3374	1.5431	1.8107	2.1509	2.5775
вариант 2	y	0	0.2013	0.4108	0.6367	0.8881	1.1752	1.5095	1.9043	2.3756
…	y	…	…	…	…	…	…	…	…	…

(где последние цифры являются продуктами правильного округления), и заданы контрольные значения аргумента

$$\bar{x}=0.25,\quad \tilde{x}=0.92\quad \text{и}\quad \hat{x}=1.63.$$

А) Записать подходящее для приближенного вычисления значений $\bar{y}=f(\bar{x})$, $\tilde{y}=f(\tilde{x})$, $\hat{y}=f(\hat{x})$ конкретные интерполяционные многочлены Лагранжа первой и второй степени и получить эти значения.

Б) Составить алгоритм, реализующий схему Эйткена вычисления с максимально возможной точностью значения $y=f(x)$ в произвольной точке x промежутка $[x_0, x_n+(x_n-x_{n-1})]$. Пользуясь этим алгоритмом, вычислить приближенные значения $\bar{y}$, $\tilde{y}$ и $\hat{y}$.

В) Составить таблицу конечных разностей, записать оптимальные для вычисления $\bar{y}$, $\tilde{y}$ и $\hat{y}$ конкретные конечноразностные формулы и с их помощью получить эти значения.

Проанализировать результаты выполнения заданий А–В.

Лабораторная работа 10. **«Аппроксимация таблично заданных функций»**

Функция $y=f(x)$ задана следующей таблицей значений

	x	10	20	30	40	50	60	70	80
вариант 1	y	2.5	3.2	3.7	4.0	4.2	4.4	4.6	4.75
вариант 2	y	4.7	6.7	8.1	9.4	10.6	11.2	12.6	13.4
…	y	…	…	…	…	…	…	…	…

1. Методом наименьших квадратов аппроксимировать $y=f(x)$:

а) линейной функцией;

б) многочленами Фурье второй, третьей и четвертой степеней;

в) функцией вида $a\lg(bx)$;

г) функцией вида ax^b;

д) функцией вида ae^{bx}.

Сравнить величины среднеквадратических отклонений. Пользуясь каждой из найденных функций а–д, вычислить контрольное приближенное значение $f(35)$.

2. А) Для функции $y=f(x)$ построить интерполяционный кубический сплайн дефекта 1 и с его помощью вычислить приближенно $f(35)$, $f'(35)$ и $\int_{10}^{40} f(x)dx$.

Б) Представить построенный сплайн линейной комбинацией кубических B-сплайнов (конкретизировать вид этих B-сплайнов соответственно рассматриваемому случаю и найти коэффициенты их линейной комбинации).

Лабораторная работа 11. **«Численное интегрирование»**

Даны интегралы:

вариант 1 $\quad I_1 := \int\limits_{0.4}^{2} \frac{1}{x} e^{0.03x} dx, \qquad I_2 := \int\limits_{0}^{4} \frac{\ln(4+x) - \ln(8)}{(4-x)\sqrt{x}} dx;$

вариант 2 $\quad I_1 := \int\limits_{0.6}^{3} \frac{1}{x + \sin 0.5x} dx, \qquad I_2 := \int\limits_{0}^{2} \frac{\ln(2-x)}{\sqrt{x}\sqrt{2-x}} dx;$

и т.д.

1. Сколько достаточно взять узлов, чтобы найти значение интеграла I_1 с точностью $\varepsilon_0 = 10^{-3}$:

а) по формулам прямоугольников (средней точки) ?

б) по формулам трапеций ?

Вычислить I_1 по этим формулам с данной точностью.

2. Вычислить I_1 по формуле Симпсона с девятью узлами и по формуле Гаусса с четырьмя узлами.

3. С точностью $\varepsilon_1 = 10^{-8}$ найти значение интеграла I_1 алгоритмом Ромберга, стараясь минимизировать количество вычислений подынтегральной функции.

Сравнить результаты пунктов 1–3.

4. Подбирая подходящие методы вычисления определенных и несобственных интегралов, с точностью $\varepsilon_2 = 10^{-4}$ найти значение I_2 не менее, чем двумя способами.

Лабораторная работа 12. **«Численное дифференцирование»**

Бесконечно гладкая функция $y = f(x)$ задана несколькими своими округленными значениями:

	x	0.2	0.3	0.4	0.5	0.6	0.7	0.8
вариант 1	y	1.3694	1.2661	1.1593	1.0472	0.9273	0.7954	0.6435
вариант 2	y	0.3948	0.5830	0.7610	0.9272	1.0808	1.2214	1.3494
...	y	...	...	...	...	...	...	...

1. Создать аналогичную таблицу с приближенными значениями функции $f'(x)$, находимыми по формулам:

а) первого порядка точности,
б) второго порядка точности,

оставляя в результатах верные цифры и один запасной десятичный знак.

2. Создать таблицу приближенных значений функции $f''(x)$, подсчитываемых по формуле второго порядка точности.

3. Максимально точно, насколько это можно в данных условиях, вычислить значения

$$y'(0.25), \quad y'(0.55), \quad y''(0.25) \quad \text{и} \quad y''(0.55).$$

Лабораторная работа 13. **«Численное решение дифференциальных уравнений первого порядка»**

Дано дифференциальное уравнение и начальное условие:

вариант 1 $\quad y' = \frac{3x^2 \cos(y^2 - x^3)}{2\sqrt{1+x^3}\cos(1)}, \qquad y(0) = 1;$

вариант 2 $\quad y' = \frac{2xe^{xy}}{(1+x^2)e^{1+x}}, \qquad y(0) = 0;$

и т.д.

1. Заполнить таблицу

x	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
y											

приближенными значениями решения данной задачи Коши, вычисленными с точностью $\varepsilon = 10^{-8}$ методом Рунге–Кутты с автоматическим выбором шага или методом Кутты–Мерсона (указать окончательный расчетный шаг в каждой точке таблицы).

2. Взяв из таблицы п.1 первые четыре значения решения, продолжить вычисления до точки $x = 1$ с фиксированным шагом $h = 0.1$ методом Милна и предиктор-корректорным методом Адамса четвертого порядка. Подсчитать главные части получаемых при этом на каждом шаге погрешностей.

Лабораторная работа 14. **«Численное решение задачи Коши для обыкновенного дифференциального уравнения второго порядка»**

Дана задача Коши:

вариант 1 $\quad y'' = -\frac{\sqrt{x+y^2}}{4\sqrt{2}x^2}, \quad y(1) = 1, \quad y'(1) = 0.5;$

вариант 2 $\quad y'' = -\frac{\sqrt{y^2 - x}}{2\sqrt{3}x^2}, \quad y(1) = 2, \quad y'(1) = 1;$

и т.д.

На отрезке $[1, 2]$ построить таблицу значений ее решения $y(x)$ с шагом $h = 0.1$ и заданной точностью $\varepsilon = 10^{-6}$, применяя:

а) сведение к системе дифференциальных уравнений первого порядка с последующим численным интегрированием ее методом Рунге–Кутты или Кутты–Мерсона:

б) предиктор-корректорные методы Адамса непосредственно к данной задаче (несколько первых «разгонных» значений можно взять из промежуточных результатов п.а)).

Указать окончательный расчетный шаг, обеспечивающий заданную точность в каждом случае.

Лабораторная работа 15. **«Численное решение линейной краевой задачи»**

Дана краевая задача:

вариант 1 $\quad y'' - \dfrac{y'}{x} - \dfrac{3y}{x^2} = \dfrac{3}{x^2}, \quad y(0.7) + 0.7y'(0.7) = -1, \quad y(1) = 0;$

вариант 2 $\quad y'' + \dfrac{y'}{x+2} + \dfrac{(4x+7)y}{4(x+2)^2} = \dfrac{1}{\sqrt{x+1}}, \quad y(2) = 2, \quad y(2.3) = 8.6y'(2.3);$

и т.д.

На промежутке, определяемом данными краевыми условиями:

1. с точностью $\varepsilon = 10^{-6}$ построить каркас решения $y(x)$ на сетке с шагом $h_0 = 0.1$ конечноразностным методом второго порядка и противопотоковым методом (указать шаг расчетной сетки, при котором обеспечивается эта точность в каждом методе);

2. применить методы Галёркина и коллокации с тремя–четырьмя базисными функциями;

3. решить краевую задачу путём сведения ее к задаче Коши (методом редукции или дифференциальной прогонки).

Результаты пп. 2, 3 сравнить с результатами п. 1 (дать сводную таблицу значений приближенных решений на сетке с шагом h_0).

Лабораторная работа 16. **«Квадратурный метод решения интегральных уравнений»**

Для уравнений

$$x(t) = \int_0^2 Q(t, s)x(s)ds + f(t) \qquad (1)$$

и

$$\int_1^t K(t, s)x(s)ds = F(t) \qquad (2)$$

заданы ядра и свободные члены:

вариант 1 $\quad Q(t, s) = 2\ln\dfrac{1+s}{1+t^2}, \quad f(t) = t^2 - t + 1;$

$K(t, s) = t + \sqrt{s}, \quad F(t) = 2t\sqrt{t} - t - 1;$

вариант 2 $\quad Q(t, s) = t + \ln(1+s), \quad f(t) = 1 - \dfrac{t^2}{t+1};$

$K(t, s) = \dfrac{\sqrt{s} - t}{s}, \quad F(t) = 3t - 2t\sqrt{t} - 1;$

и т.д.

1. На сетке точек t_i отрезка $[0, 2]$ с шагом сетки $h_1 = 0.5$ построить каркас приближенного решения уравнения (1) с точностью $\varepsilon = 10^{-6}$, пользуясь какой-либо квадратурной формулой замкнутого типа и применяя сгущающиеся расчетные сетки для обеспечения заданной точности. На основе полученного каркаса записать приближенное решение в виде непрерывной функции и с ее помощью вычислить приближенные значения $x\left(\dfrac{1}{e}\right)$ и $x\left(\dfrac{\pi}{2}\right)$.

2. Применяя квадратурную формулу прямоугольников на отрезке $[1, 2]$ с шагом $h_2 = 0.2$, найти каркас приближенного решения уравнения (2) с точностью $\varepsilon = 10^{-4}$. Представить полученное дискретное решение интерполяционным многочленом третьей степени, построенным по первым четырем узлам заданной сетки, и, пользуясь этим, вычислить приближенно $x\left(\dfrac{e^2}{5}\right)$ и $x\left(\dfrac{\pi^2}{9}\right)$.

Лабораторная работа 17. **«Численное решение смешанной задачи для параболического уравнения методом конечных разностей»**

Дана начально-граничная задача для неоднородного уравнения теплопроводности:

вариант 1 $\quad \dfrac{\partial u}{\partial t} = 4\dfrac{\partial^2 u}{\partial x^2} + \ln\dfrac{x+1}{t+1}, \quad x \in [0, 1], \quad t \in [0, T];$

$u(x, 0) = \dfrac{1}{x+1}; \quad u(0, t) = \cos t, \quad u(1, t) = 0.5; \quad T = 0.3;$

вариант 2 $\dfrac{\partial u}{\partial t}=0.25\dfrac{\partial^2 u}{\partial x^2}+e^{-xt}$, $x\in[0,1]$, $t\in[0,T]$;

$u(x,0)=\operatorname{arctg}x$; $u(0,t)=0$, $u(1,t)=\left(\dfrac{\pi}{4}-t\right)\cos t$; $T=2$;

и т.д.

Решить ее методом конечных разностей, обеспечивая точность $\varepsilon=0.01$ в узлах сетки с шагом $h=0.2$ с помощью правила Рунге. Сравнить требуемый для этого объем вычислений при использовании следующих двухслойных разностных схем:

а) явной четырехточечной;
б) неявной четырехточечной;
в) схемы Кранка–Николсон.

Лабораторная работа 18. **«Конечноразностное решение смешанной задачи для гиперболического уравнения»**

Дана задача:

вариант 1 $4\dfrac{\partial^2 u}{\partial t^2}=\dfrac{\partial^2 u}{\partial x^2}+e^{xt}$, $x\in[0,1]$, $t\in[0,T]$;

$u(x,0)=x$; $\dfrac{\partial u(x,0)}{\partial t}=\sqrt{x}$, $u(0,t)=\operatorname{arctg}t$, $u(1,t)=1$;

вариант 2 $\dfrac{\partial^2 u}{\partial t^2}=4\dfrac{\partial^2 u}{\partial x^2}+\ln\left(1+xt^2\right)$, $x\in[0,1]$, $t\in[0,T]$;

$u(x,0)=\sqrt{1-x^2}$, $\dfrac{\partial u(x,0)}{\partial t}=x^2$, $u(0,t)=1$, $u(1,t)=\sin t$;

и т.д.

Для значения $T=0.5$ получить ее численное решение на сетке с шагом $h=0.2$ по переменной x, применяя явную разностную схему второго порядка. Для вычисления значений первого слоя использовать как простейшую несимметричную аппроксимацию $\dfrac{\partial u}{\partial t}$ в начальном условии, так и метод фиктивной точки.

Лабораторная работа 19. **«Метод переменных направлений решения двумерной задачи теплопроводности»**

Применяя продольно-поперечную прогонку, найти каркас приближенного решения задачи

$\dfrac{\partial u}{\partial t}=\dfrac{\partial^2 u}{\partial x^2}+\dfrac{\partial^2 u}{\partial y^2}+f(x,y,t)$, $(x,y,t)\in\Omega\times[0,T]$;

$u(x,y,0)=\varphi(x,y)$, $(x;y)\in\Omega$;

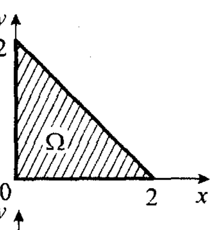
$u(x,y,t)=\psi(x,y,t)$, $(x;y)\in\Gamma_\Omega$

(где область Ω и функции f,φ,ψ определены ниже) **на квадратной сетке** с шагом $h=0.25$ при $T=0.1$. Провести расчеты со следующими значениями шага τ по времени t:

а) $\tau:=T$; б) $\tau:=\dfrac{T}{2}$; в) $\tau:=\dfrac{T}{4}$; г) $\tau:=\dfrac{T}{8}$.

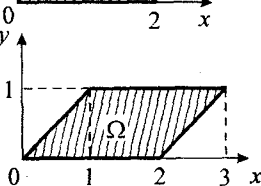
вариант 1

$f(x,y,t)=x^2+yt$,
$\varphi(x,y)=1$,
$\psi(x,y,t)=e^{-xyt}$;



вариант 2

$f(x,y,t)=(1+t^2)\sin\pi y$,
$\varphi(x,y)=0$,
$\psi(x,y,t)=\ln(1+xyt)$;



и т.д.

Лабораторная работа 20. **«Конечноразностное решение задачи Дирихле для уравнения Пуассона»**

Получить каркас приближенного решения заданной ниже граничной задачи на квадратной сетке с шагом $h=0.2$ с точностью $\varepsilon=0.01$ (расчетный шаг устанавливается по правилу Рунге). Применить:

1) метод конечных разностей с решением систем сеточных уравнений: а) методом Гаусса; б) методами Зейделя и ПВР (с эмпирической оптимизацией параметра релаксации);

2) итерационный метод переменных направлений (с нулевым начальным приближением).

вариант 1 $\dfrac{\partial^2 u}{\partial x^2}+\dfrac{\partial^2 u}{\partial y^2}-0.1y\dfrac{\partial u}{\partial x}+x^2=0$, $(x;y)\in\Omega$,

$\Omega:=\{(x;y)|\ \ x>y,\ \ x+y<2,\ \ y>0\}$,

$u|_{\Gamma_\Omega}=0$;

вариант 2 $\dfrac{\partial^2 u}{\partial x^2}+\dfrac{\partial^2 u}{\partial y^2}-0.2x\dfrac{\partial u}{\partial y}+\sqrt{y}=0$, $(x;y)\in\Omega$,

$\Omega:=\{(x;y)|\ \ y>x,\ \ x+y<2,\ \ x>0\}$,

$u|_{\Gamma_\Omega}=0$;

и т.д.

ЛИТЕРАТУРА

1. *Алберг Дж., Нильсон Э., Уолш Дж.* Теория сплайнов и ее приложения. — М.: Мир, 1972.

2. Алгоритмы, математическое обеспечение и архитектура многопроцессорных вычислительных систем / Отв. ред. *А.П.Ершов*. — М.: Наука, 1982.

3. *Амосов А.А., Дубинский Ю.А., Копчёнова Н.В.* Вычислительные методы для инженеров. — М.: Высшая школа, 1994.

4. *Андреев В.Б., Руховец Л.А.* Проекционные методы // Математика, кибернетика. — №11. — М.: Знание, 1986.

5. *Араманович И.Г., Левин В.И.* Уравнения математической физики. — М.: Наука, 1969.

6. *Арушанян О.Б., Залеткин С.В.* Численное решение обыкновенных дифференциальных уравнений на Фортране. — М.: Изд-во МГУ, 1990.

7. *Ахиезер Н.И.* Лекции по теории аппроксимации. — М.: Наука, 1965.

8. *Ахромеева Т.С., Курдюмов С.П., Малинецкий Г.Г.* Парадоксы мира нестационарных структур. В кн. «Компьютеры и нелинейные явления». — М.: Наука, 1988.

9. *Бабенко К.И.* Основы численного анализа. — М: Наука, 1986.

10. *Бабушка И., Витасек Э., Прагер М.* Численные процессы решения дифференциальных уравнений. — М.: Мир, 1969.

11. *Бакушинский А.Б., Гончарский А.В.* Итеративные методы решения некорректных задач. — М.: Наука, 1989.

12. *Бахвалов Н.С.* Численные методы. — М.: Наука, 1973.

13. *Бахвалов Н.С., Жидков Н.П., Кобельков Г.М.* Численные методы. — М.: Наука, 1987.

14. *Бахвалов Н.С., Жидков Н.П., Кобельков Г.М.* Численные методы. — М.: Лаборатория Базовых Знаний, 2001.

15. *Бахвалов Н.С., Лапин А.В., Чижонков Е.В.* Численные методы в задачах и упражнениях. — М.: Высшая школа, 2000.

16. *Бейкер Дж., Грейвс–Моррис П.* Аппроксимации Паде. — М.: Мир, 1986.

17. *Беланов А.А.* Решение алгебраических уравнений методом Лобачевского. — М.: Наука, 1989.

18. *Бердышев В.И., Петрак Л.В.* Аппроксимация функций, сжатие численной информации, приложения. — Екатеринбург: УрО РАН, 1999.

19. *Березин И.С., Жидков Н.П.* Методы вычислений. Т.1. — М.: Физматгиз, 1962.

20. *Березин И.С., Жидков Н.П.* Методы вычислений. Т.2. — М.: Физматгиз, 1962.

21. *Берс Л., Джон Ф., Шехтер М.* Уравнения с частными производными. — М.: Мир, 1986.

22. *Биргер И.А., Шорр Б.Ф., Шнейдерович Р.М.* Расчет на прочность деталей машин. — М.: Машиностроение, 1966.

23. *Боглаев Ю.П.* Вычислительная математика и программирование. — М.: Высшая школа, 1990.

24. *Бор К.* Практическое руководство по сплайнам. — М.: Радио и связь, 1985.

25. *Брадис В.М.* Вычислительная работа в курсе математики средней школы. — М.: Изд.АПН РСФСР, 1962.

26. *Бут Э.Д.* Численные методы. — М.: Физматгиз, 1959.

27. *Вазов В., Форсайт Дж.* Разностные методы решения дифференциальных уравнений в частных производных. — М.: ИЛ, 1963.

28. *Вайникко Г.М., Веретенников А.Ю.* Итерационные процедуры в некорректных задачах. — М.: Наука, 1986.

29. *Валях Е.* Последовательно-параллельные вычисления. — М.: Мир, 1985.

30. *Вапник В.Н.* Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979.

31. *Варга Р.* Функциональный анализ и теория аппроксимации в численном анализе. — М.: Мир, 1974.

32. *Васильев Ф.П.* Численные методы решения экстремальных задач. — М.: Наука, 1988.

33. *Васин В.В., Агеев А.Л.* Катастрофы и парадоксы при решении неустойчивых задач на ЭВМ // Математика, кибернетика. — №10. — М.: Знание, 1991.

34. *Васин В.В., Агеев А.Л.* Некорректные задачи с априорной информацией. — Екатеринбург: УИФ «Наука», 1993.

35. *Вержбицкий В.М.* Выбор параметров в теоремах сходимости одного аппроксимационного аналога метода Ньютона // Журн. вычислит. математики и мат. физики, 1975, **15**. — № 6, с.1594–1597.

36. *Вержбицкий В.М.* Обращение матриц и решение нелинейных систем. — Ижевск: Изд. ИМИ, 1980.

37. *Вержбицкий В.М.* Численные методы (линейная алгебра и нелинейные уравнения). — М.: Высшая школа, 2000.

38. *Вержбицкий В.М.* Численные методы (математический анализ и обыкновенные дифференциальные уравнения). — М.: Высшая школа, 2001.

39. *Вержбицкий В.М., Цалюк З.Б.* Об усиленном методе Ньютона–Канторовича с аппроксимацией обратного оператора // Журн. вычислит. математики и мат. физики, 1972, **12**. — №1, с.222–227.

40. *Вержбицкий В.М., Цалюк З.Б.* Об одном аналоге усиленного метода Ньютона–Канторовича // Докл. АН СССР, 1972, **203**. — №3, с.515–516.

41. *Верлань А.Ф., Сизиков В.С.* Интегральные уравнения: методы, алгоритмы, программы. — Киев: Наукова думка, 1986.

42. *Воеводин В.В.* Вычислительные основы линейной алгебры. — М.: Наука, 1977.

43. *Воеводин В.В., Кузнецов Ю.А.* Матрицы и вычисления. — М.: Наука, 1984.

44. *Волков Б.А.* Численные методы. — М.: Наука, 1979.

45. *Вулих Б.З.* Введение в функциональный анализ. — М.: Наука, 1967.

46. Вычислительная математика / *Н.И.Данилина, Н.С.Дубровская, О.П.Кваша и др.* — М.: Наука, 1985.

47. *Гаврурин М.К.* Лекции по методам вычислений. — М.: Наука, 1971.

48. *Гаврурин М.К.* Нелинейные функциональные уравнения и непрерывные аналоги итерационных методов // Изв. вузов: Математика, 1958. — № 5(6), с.18–31.

49. *Галеев Э.М., Тихомиров В.Н.* Оптимизация: теория, примеры, задачи. — М.: Эдиториал УРСС, 2000.

50. Гарантированная точность решения систем линейных уравнений в евклидовых пространствах / *С.К.Годунов, А.Г.Антонов и др.* — Новосибирск: ВО «Наука», 1992.

51. *Гельфонд А.О.* Исчисление конечных разностей. — М.: Физматгиз, 1959.

52. *Гилл Ф., Мюррей У., Райт М.* Практическая оптимизация. — М.: Мир, 1985.

53. *Годунов С.К.* Уравнения математической физики. — М.: Наука, 1979.

54. *Годунов С.К.* Решение систем линейных уравнений. — Новосибирск: Наука, 1980.

55. *Годунов С.К., Рябенький В.С.* Введение в теорию разностных схем. — М.: Физматгиз, 1962.

56. *Годунов С.К., Рябенький В.С.* Разностные схемы (введение в теорию). — М.: Наука, 1973.

57. *Гончаров В.Л.* Теория интерполирования и приближения функций. Изд. 2, перераб.— М.: ГТТИ, 1954.

58. *Гутер Р.С., Овчинский Б.В.* Элементы численного анализа и математической обработки результатов опыта. — М.: Наука, 1970.

59. *Двайт Г.Б.* Таблицы интегралов и другие математические формулы. — М.: Наука, 1966.

60. *Деккер К., Вервер Я.* Устойчивость методов Рунге-Кутты для жестких нелинейных дифференциальных уравнений. — М.: Мир, 1988.

61. *Демидович Б.П., Марон И.А.* Основы вычислительной математики. — М.: Наука, 1970.

62. *Демидович Б.П., Марон И.А.,Шувалова Э.З.* Численные методы анализа. — М.: Наука, 1967.

63. *Демьянов В.Ф., Малоземов В.Н.* Введение в минимакс. — М.: Наука, 1972.

64. *Джордж А., Лю Дж.* Численное решение больших разреженных систем уравнений. — М.: Мир, 1984.

65. *Дробышевич В.И., Дымников В.П., Ривин Г.С.* Задачи по вычислительной математике. — М.: Наука, 1980.

66. *Дьяконов Е.Г.* Минимизация вычислительной работы. Асимптотически оптимальные алгоритмы для эллиптических задач. — М.: Наука, 1989.

67. *Дьяченко В.Ф.* Основные понятия вычислительной математики. — М.: — Наука, 1972.

68. *Дэннис Дж., Шнабель Р.* Численные методы безусловной оптимизации и решения нелинейных уравнений. — М.: Мир, 1988.

69. *Ермаков С.М.* Метод Монте-Карло и смежные вопросы. — М.: Наука, 1971.

70. *Ефимов А.В., Золотарев Ю.Г., Терпигорева В.М.* Математический анализ (специальные разделы). Ч. II. Применение некоторых методов математического и функционального анализа. — М.: Высш. школа, 1980.

71. *Журкин И.Г., Нейман Ю.М.* Методы вычислений в геодезии. — М.: Недра, 1988.

72. *Загускин В.Л.* Справочник по численным методам решения алгебраических и трансцендентных уравнений. — М.: Физматгиз, 1960.

73. *Иванов В.В.* Методы вычислений на ЭВМ: Справочное пособие. — Киев: Наукова думка, 1986.

74. *Икрамов Х.Д.* Численные методы линейной алгебры (решение линейных уравнений) // Математика, кибернетика. — №4. — М.: Знание, 1987.

75. *Икрамов Х.Д.* Несимметричная проблема собственных значений. — М.: Наука, 1991.

76. *Ильин В.П., Кузнецов Ю.А.* Трехдиагональные матрицы и их приложения. — М.: Наука, 1985.

77. Интегральные уравнения / *П.П.Забрейко, А.И.Кошелев, М.А. Красносельский и др.* — М.: Наука, 1968.

78. *Калиткин Н.Н.* Численные методы. — М.: Наука, 1978.

79. *Камке Э.* Справочник по обыкновенным дифференциальным уравнениям. — М.: Наука, 1976.

80. *Канторович Л.В., Акилов Г.П.* Функциональный анализ. — М.: Наука, 1977.

81. *Канторович Л.В., Крылов В.И.* Приближенные методы высшего анализа. — М.–Л.: Физматгиз, 1962.

82. *Каханер Д., Моулер К., Нэш С.* Численные методы и программное обеспечение. — М.: Мир, 1998.

83. *Коваленко И.Н., Филиппова А.А.* Теория вероятностей и математическая статистика. — М.: Высшая школа, 1973.

84. *Коллатц Л.* Численные методы решения дифференциальных уравнений. — М.: ИЛ, 1953.

85. *Коллатц Л.* Задачи на собственные значения с техническими приложениями. — М. : Наука, 1968.

86. *Коллатц Л.* Функциональный анализ и вычислительная математика. — М.: Мир, 1969.

87. *Коллатц Л., Альбрехт Ю.* Задачи по прикладной математике. — М.: Мир, 1978.

88. *Коллатц Л., Крабс В.* Теория приближений. Чебышевские приближения и их приложения. — М.: Наука, 1978.

89. *Копченова Н.В., Марон И.А.* Вычислительная математика в примерах и задачах. — М.: Наука, 1972.

90. *Корн Г., Корн Т.* Справочник по математике для научных работников и инженеров. — М.: Наука, 1968.

91. *Корнейчук Н.П.* Сплайны в теории приближений. — М.: Наука, 1984.

92. *Косарев В.И.* 12 лекций по вычислительной математике. — М.: Изд-во МФТИ, 1995.

93. *Кочиков И.В., Курамшина Г.М., Ягола А.Г.* Численные методы в колебательной спектроскопии // Математика, кибернетика. — №1. — М.: Знание, 1989.

94. *Краскевич В.Е., Зеленский К.Х., Гречко В.И.* Численные методы в инженерных исследованиях. — Киев: Вища школа, 1986.

95. *Краснов М.Л.* Интегральные уравнения. — М.: Наука, 1975.

96. *Краснов М.Л., Киселев А.И., Макаренко Г.И.* Интегральные уравнения. Задачи и упражнения. — М.: Наука, 1968.

97. *Кроновер Р.М.* Фракталы и хаос в динамических системах. Основы теории. — М.: Постмаркет, 2000.

98. *Крылов А.Н.* Лекции о приближенных вычислениях. Изд. 5. — М.–Л.: ГТТИ, 1950.

99. *Крылов В.И., Бобков В.В., Монастырный П.И.* Вычислительные методы. Т.1. — М.: Наука, 1976.

100. *Крылов В.И., Бобков В.В., Монастырный П.И.* Вычислительные методы. Т.2. — М.: Наука, 1977.

101. *Крылов В.И., Бобков В.В., Монастырный П.И.* Начала теории вычислительных методов. Интерполирование и интегрирование. — Минск: Наука и техника, 1983.

102. *Крылов В.И., Бобков В.В., Монастырный П.И.* Начала теории вычислительных методов. Линейная алгебра и нелинейные уравнения. — Минск: Наука и техника, 1985.

103. *Крылов В.И., Бобков В.В., Монастырный П.И.* Начала теории вычислительных методов. Уравнения в частных производных. — Минск: Наука и техника, 1986.

104. *Крылов В.И., Шульгина Л.Т.* Справочная книга по численному интегрированию. — М.: Наука, 1966.

105. *Курант Р.* Курс дифференциального и интегрального исчисления. Т.1. — М.: Наука, 1967.

106. *Ланс Дж.Н.* Численные методы для быстродействующих вычислительных машин. — М.: ИЛ, 1962.

107. *Ланцош К.* Практические методы прикладного анализа. — М.: Физматгиз, 1961.

108. *Лесин В.В., Лисовец Ю.П.* Основы методов оптимизации. — М.: Изд-во МАИ, 1995.

109. *Лизоркин П.И.* Курс дифференциальных и интегральных уравнений с дополнительными главами анализа. — М.: Наука, 1981.

110. *Ловитт У.В.* Линейные интегральные уравнения. — М.: ГИТТЛ, 1957.

111. *Локуциевский О.В., Гавриков М.Б.* Начала численного анализа. — М.: ТОО «Янус», 1995.

112. *Люстерник Л.А., Червоненкис О.А., Янпольский А.Р.* Математический анализ. Вычисление элементарных функций. — М.: Физматгиз, 1963.

113. *Магомедов К.М., Холодов А.С.* Сеточно–характеристические численные методы. — М.: Наука, 1988.

114. *Макаров В.Л., Хлобыстов В.В.* Сплайн–аппроксимация функций. — М.: Высшая школа, 1983.

115. *Мак–Кракен Д., Дорн У.* Численные методы и программирование на Фортране. — М.: Мир, 1977.

116. *Манжиров А.В., Полянин А.Д.* Справочник по интегральным уравнениям: Методы решения. — М.: Факториал Пресс, 2000.

117. *Марчук Г.И.* Методы вычислительной математики. — М.: Наука, 1977.

118. *Марчук Г.И.* Методы расщепления. — М.: Наука, 1988.

119. Математика и САПР. В 2-х кн. Кн. 1 / *П.Шенен, М.Коснар, И.Гардан и др.* — М.: Мир, 1988.

120. Математика и САПР: В 2–х кн. Кн. 2 / *П.Жермен–Лакур, П.Л.Жорж, Ф.Пистр и др.* — М.: Мир, 1989.

121. Математическая статистика / *В.М.Иванова, В.Н.Калинина и др.* — М.: Высшая школа, 1975.

122. Метод статистических испытаний (метод Монте-Карло) / *Н.П.Бусленко, Д.И.Гуленко, И.М.Соболь и др.* — М.: Физматлит, 1962.

123. *Микеладзе Ш.Е.* Численные методы математического анализа. — М.: ГТТИ, 1953.

124. *Митчел Э., Уэйт Р.* Метод конечных элементов для уравнений с частными производными. — М.: Мир, 1981.

125. *Михлин С.Г.* Численная реализация вариационных методов. — М.: Наука, 1966.

126. *Михлин С.Г., Смолицкий Х.Л.* Приближенные методы решения дифференциальных и интегральных уравнений. — М.: Наука, 1965.

127. *Молчанов И.Н.* Машинные методы решения прикладных задач. Дифференциальные уравнения. — Киев: Наукова думка, 1988.

128. *Морозов В.А.* Регулярные методы решения некорректно поставленных задач. — М.: Наука, 1987.

129. *Мысовских И.П.* Лекции по методам вычислений. — М.: Физматгиз, 1962.

130. *Мышкис А.Д.* Математика для втузов. Специальные курсы. — М.: Наука, 1971.

131. *На Цунг–Йен.* Вычислительные методы решения прикладных задач. Дифференциальные уравнения. — М.: Мир, 1982.

132. *Никифоров А.Ф., Суслов С.К.* Классические ортогональные полиномы // Математика, кибернетика. — №12. — М.: Знание, 1985.

133. *Никольский С.М.* Квадратурные формулы. — М.: Наука, 1988.

134. *Носач В.В.* Решение задач аппроксимации с помощью персональных компьютеров. — М.: МИКАП, 1994.

135. *Норри Д., де Фриз Ж.* Введение в метод конечных элементов. — М.: Мир, 1981.

136. *Обэн Ж.-П.* Приближенное решение эллиптических краевых задач. — М.: Мир, 1977.

137. *Ортега Дж.* Введение в параллельные и векторные методы решения линейных систем. — М.: Мир, 1991.

138. *Ортега Дж., Пул У.* Введение в численные методы решения дифференциальных уравнений. — М.: Наука, 1986.

139. *Ортега Дж., Рейнболдт В.* Итерационные методы решения нелинейных систем уравнений со многими неизвестными. — М.: Мир, 1975.

140. *Островский А.М.* Решение уравнений и систем уравнений. — М.: ИЛ, 1963.

141. *Парлетт Б.* Симметричная проблема собственных значений. — М.: Мир, 1983.

142. *Плис А.И., Сливина Н.А.* Лабораторный практикум по высшей математике. — М.: Высшая школа, 1994.

143. *Победря Б.Е.* Численные методы в теории упругости и пластичности. — М.: Изд-во МГУ, 1995.

144. *Полянин А.Д.* Справочник по линейным уравнениям математической физики. — М.: Физматлит, 2001.

145. *Попов Б.А., Теслер Г.С.* Приближение функций для технических приложений. — Киев: Наукова думка, 1980.

146. *Прангишвили И.В., Виленкин С.Я., Медведев И.Л.* Параллельные вычислительные системы с общим управлением. — М.: Энергоатомиздат, 1983.

147. Приближенное решение операторных уравнений / *М.А.Красносельский, Г.М.Вайникко, П.П.Забрейко и др.* — М.: Наука, 1969.

148. *Пугачев В.С.* Лекции по функциональному анализу. — М.: Изд-во МАИ, 1996.

149. *Ракитин В.И., Первушин В.Е.* Практическое руководство по методам вычислений с приложением программ для персональных компьютеров. — М.: Высшая школа, 1998.

150. *Ремез Е.Я.* Основы численных методов чебышевского приближения. — Киев: Наукова думка, 1969.

151. *Рихтмайер Р., Мортон К.* Разностные методы решения краевых задач. — М.: Мир, 1972.

152. *Рябенький В.С.* Введение в вычислительную математику. — М.: Наука, 1994.

153. *Самарский А.А.* Введение в численные методы. — М.: Наука, 1987.

154. *Самарский А.А.* Теория разностных схем. — М.: Наука, 1989.

155. *Самарский А.А., Андреев В.Б.* Разностные методы для эллиптических уравнений. — М.: Наука, 1976.

156. *Самарский А.А., Вабищевич П.Н.* Численные методы решения задач конвекции–диффузии. — М.: Эдиториал УРСС, 1999.

157. *Самарский А.А., Гулин А.В.* Устойчивость разностных схем. — М.: Наука, 1973.

158. *Самарский А.А., Гулин А.В.* Численные методы. — М.: Наука, 1989.

159. *Самарский А.А., Гулин А.В.* Численные методы математической физики. — М.: Научный мир, 2000.

160. *Самарский А.А., Лазаров Р.Д., Макаров В.Л.* Разностные схемы для дифференциальных уравнений с обобщенными решениями. — М.: Высшая школа, 1987.

161. *Самарский А.А., Николаев Е.С.* Методы решения сеточных уравнений. — М.: Наука , 1978.

162. Сборник задач по математике для втузов. Ч.4. Методы оптимизации. Уравнения в частных производных. Интегральные уравнения / *Э.А.Вуколов, А.В.Ефимов, В.Н.Земсков и др.* Под ред. *А.В.Ефимова* — М.: Наука, 1990.

163. Сборник задач по методам вычислений / Под ред. *П.И.Монастырного* — М.: Наука, 1994.

164. *Смирнов В.И.* Курс высшей математики. Т.2. — М.: Наука, 1967.

165. *Смирнов В.И.* Курс высшей математики. Т.4., Ч.1. — М.: Наука, 1974.

166. *Степанов В.В.* Курс дифференциальных уравнений. Изд.6. — М.: ГИТТЛ, 1953.

167. *Стечкин С.Б., Субботин Ю.Н.* Сплайны в вычислительной математике. — М.: Наука, 1976.

168. *Стренг Г., Фикс Дж.* Теория метода конечных элементов. — М.: Мир, 1977.

169. *Суетин П.К.* Классические ортогональные многочлены. — М.: Наука, 1976.

170. *Сьярле Ф.* Метод конечных элементов для эллиптических задач. — М.: Мир, 1980.

171. *Талдыкин А.Т.* Элементы прикладного функционального анализа. — М.: Высшая школа, 1982.

172. Теоретические основы конструирования численных алгоритмов задач математической физики / *Н.Н.Анучина, К.И.Бабенко, С.К.Годунов и др.* — М.: Наука, 1979.

173. *Тиман А.Ф.* Теория приближения функций действительного переменного. — М.: Физматгиз, 1960.

174. *Тихонов А.Н., Костомаров Д.П.* Вводные лекции по прикладной математике. — М.: Наука, 1984.

175. *Тихонов А.Н., Арсенин В.Я.* Методы решения некорректных задач. — М.: Наука, 1985.

176. *Трауб Дж.* Итерационные методы решения уравнений. — М.: Мир, 1985.

177. *Треногин В.А.* Функциональный анализ. — М.: Наука, 1980.

178. *Турчак Л.И.* Основы численных методов. — М.: Наука, 1987.

179. *Уилкинсон Дж.Х.* Алгебраическая проблема собственных значений. — М.: Наука, 1970.

180. *Фаддеев Д.К., Фаддеева В.Н.* Вычислительные методы линейной алгебры. — М.: Физматгиз, 1960.

181. *Федоренко Р.П.* Введение в вычислительную физику. — М.: Изд-во МФТИ, 1994.

182. *Флетчер К.* Численные методы на основе метода Галёркина. — М.: Мир, 1988.

183. *Форсайт Дж., Молер К.* Численное решение систем линейных алгебраических уравнений. — М.: Мир, 1969.

184. *Форсайт Дж., Малькольм М., Моулер К.* Машинные методы математических вычислений. — М.: Мир, 1980.

185. *Хайрер Э., Нёрсетт С., Ваннер Г.* Решение обыкновенных дифференциальных уравнений. Нежесткие задачи. — М.: Мир, 1990.

186. *Хейгеман Л., Янг Д.* Прикладные итерационные методы. — М.: Мир, 1986.

187. *Хемминг Р.В.* Численные методы. — М.: Наука, 1968.

188. *Холл Дж., Уатт Дж.* Современные численные методы решения обыкновенных дифференциальных уравнений. — М.: Мир, 1979.

189. *Цимринг Ш.Е.* Специальные функции и определенные интегралы. Алгоритмы. Программы для микрокалькуляторов: Справочник. — М.: Радио и связь, 1988.

190. *Цлаф Л.Я.* Вариационное исчисление и интегральные уравнения. —М.: Наука, 1970.

191. *Шайдуров В.В.* Многосеточные методы конечных элементов. — М.: Наука, 1989.

192. *Шаманский В.Е.* Методы численного решения краевых задач на ЭЦВМ. — Киев: Наукова думка. 1966.

193. *Шарковский А.Н., Майстренко Ю.Л., Романенко Е.Ю.* Разностные уравнения и их приложения. — Киев: Наукова думка, 1986.

194. *Шолохович Ф.А., Васин В.В.* Основы высшей математики. — Екатеринбург: Изд-во Урал. ун-та, 1999.

195. *Штеттер X.* Анализ методов дискретизации для обыкновенных дифференциальных уравнений. — М.: Мир, 1978.

196. *Эльсгольц Л.Э.* Дифференциальные уравнения и вариационное исчисление. — М.: Наука, 1969.

197. *Эстербю О., Златев 3.* Прямые методы для разреженных матриц. — М.: Мир, 1987.

198. Математическая энциклопедия. ТТ. 1–5. — М.: Советская энциклопедия, 1977.

199. Математический энциклопедический словарь. — М.: Советская энциклопедия, 1988.

200. Словарь иностранных слов. — М.: Русский язык, 1989.

201. *Altman M.* An optimum cubically convergent iterative method of inverting a linear bounded operator in hilbert space // Pacific J.Math., 10, 1960, № 4, 1107–1113.

202. *Crank J., Nicolson P.* A practical method for numerical evalution of solution of partial differential equations of the heat-conduction type // Proc. Cambridge Philos. Soc., 43 (1947), 50–67. [Re-published in: John Crank 80th birthday special issue Adv. Comput. Math., 6 (1997), 207–226].

203. *Diaconu A.* Sur quelques méthodes itératives combinées // Matematica (RSR), 22(45), 1980, № 2, 247–261.

204. *Peaceman D.W., Rachford H.H.* The numerical solution of parabolic and elliptic differential equations // J. Indust. Appl. Math., 1955, № 3, 28–41.

205. *Schulz G.* Iterative Berechnung der reziproken Matrix // ZAMM, 13, 1933, 57–59.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

УКАЗАТЕЛЬ ОБОЗНАЧЕНИЙ И СОКРАЩЕНИЙ

$\hat{T}_n(x)$	нормированный многочлен Чебышева $T_n(x)/2^{n-1}$ 386
$\chi_n(x)$	многочлен Лежандра 420
$B_{m,k}(x)$	B-сплайн степени m с базовым узлом x_k 453
I	1) значение интеграла $\int_a^b f(x)\,dx$ 466 2) тождественный оператор 800
I^{Π}, I^T, I^C	значения интеграла, полученные по формулам прямоугольников, трапеций и Симпсона соответственно 468, 478, 480
$R(h)$	поправка Ричардсона 483, 552
$R_n[a,b]$	n-мерное пространство определенных на $[a,b]$ сеточных функций 412
$C^k[a,b]$	пространство k раз непрерывно дифференцируемых на $[a,b]$ функций 194, 336, 796
$L_2[a,b]$	пространство функций, измеримых на $[a,b]$ и интегрируемых с квадратом 638, 796, 798
H	гильбертово пространство 637, 797
$\rho_X(x_1,x_2)$	расстояние между x_1, $x_2 \in X$ (метрика) 588, 790
(x,y)	скалярное произведение элементов x и y 796
$\lVert\cdot\rVert$	норма элемента, оператора, функционала 793, 799, 802
$\lvert\!\lVert\cdot\rVert\!\rvert$	энергетическая норма 713
$[\cdot,\cdot]$	энергетическое скалярное произведение 712
$\lVert\cdot\rVert_p$	l_p-норма (Гёльдера) 109, 795
$\lVert\mathbf{A}\rVert_F$	норма Фробениуса матрицы $\mathbf{A}$ 166, 801
A^{-1}	обратный к A оператор 40, 800
A^*	сопряженный к A оператор 803
$L(X,Y)$	пространство линейных операторов из X в Y 800
у.п.x, б.п.x	условная и безусловная погрешности корня x 35, 36, 196

ЭВМ	электронно-вычислительная машина (компьютер) 22
СЛАУ	система линейных алгебраических уравнений 52, 91, 136, 732, 768
МПИ	метод простых итераций 92, 244
ПВР	последовательная верхняя релаксация 113, 778
МСГ	метод сопряженных градиентов 119
ААМН	аппроксимационный аналог метода Ньютона 289
МНК	метод наименьших квадратов 406
ОДУ	обыкновенное дифференциальное уравнение 533
МКР	метод конечных разностей 626, 727
МКЭ	метод конечных элементов 642, 722
БПФ	быстрое преобразование Фурье 775
ПТИМ	попеременно-треугольный итерационный метод 786

Учебное издание

Вержбицкий Валентин Михайлович

ОСНОВЫ ЧИСЛЕННЫХ МЕТОДОВ

Редактор *Л. В. Честная*
Художественный редактор *Ю. Э. Иванова*
Оригинал-макет выполнен *Ю.В. Гавриным*

Лицензия ИД № 06236 от 09.11.01.

Изд № ФМ-227. Подп. в печать 09.04.02. Формат $60\times88^1/_{16}$. Бум. газетн. Гарнитура «Таймс». Печать офсетная. Объем 51,94 усл. печ. л., 51,94 усл. кр.-отт., 48,26 уч.-изд. л. Тираж 6000 экз. Заказ Б-223

ФГУП «Издательство «Высшая школа», 127994, Москва, ГСП-4, Неглинная ул., 29/14.

Тел.: (095) 200-04-56. E-mail: info@v-shkola.ru http://www.v-shkola.ru

Отдел продаж: (095) 200-07-69, 200-59-39, факс: (095) 200-03-01. E-mail: sales@v-shkola.ru

Отдел «Книга-почтой»: (095) 200-33-36. E-mail: bookpost@v-shkola.ru

Отпечатано в ГУП ПИК «Идел-Пресс». 420066, г. Казань, ул. Декабристов, д. 2.

ISBN 5-06-004020-8

9 785060 040203

V. M. Verzhbitsky
Foundations of Numerical Methods
Moscow, V. Shkola, 2002, 848 pp.

Chapter 1: ***On Computational Errors***

The first chapter considers issues related to computational errors that are unavoidable in any numerical analysis of mathematical models, for example, when solving linear algebraic systems and non-linear scalar equations. It emphasizes the unavoidably approximate nature of computer operations with real numbers and provides examples of problems and methods that are extremely sensitive to errors in input data and in arithmetic operations. The chapter gives a first introduction to well- and ill-posed (by Hadamard and by Tikhonov) problems, regularization algorithms and Tikhonov's α-regularization method.

Chapter 2: ***Solution of Linear Algebraic Systems (Direct Methods)***

This chapter considers simple and frequently used numerical methods of solving sets of linear algebraic equations with square coefficient matrices, and concurrently solves the problems of inverting matrices and computing determinants. Along with the well-known Gauss method, which is studied here from the viewpoint of real-world calculations, the reader is also presented with the rotation method, which is more stable with respect to approximation errors in arithmetic operations than the Gauss method. Also considered is the square root method for solving systems with symmetrical matrices and the sweep method for systems with tridiagonal matrices (for example, for second-order three-point finite difference schemes).

Chapter 3: ***Iterative Methods for Solving Linear Algebraic Systems and Inverting Matrices***

Covered here are iterative methods of solving sets of linear algebraic equations and inverting matrices. Such methods represent a viable alternative to direct methods, at least in cases of high dimensionality. It demonstrates the logic used to construct some of the most important iterative processes such as the simple iterations method, the Jacobi method, the Seidel method, the relaxation method and the Schulz method, and analyzes the conditions for the convergence of these methods to the correct solutions. The chapter also gives a brief introduction to establishment methods, describes the algorithm of the conjugate gradient method, and explains the essence of the least-residual method.

Chapter 4: ***Methods of Solving Algebraic Eigenvalue Problems***

Chapter 4 touches upon the most difficult problem of computational linear algebra, i.e. finding the eigenvalues and eigenvectors of a matrix. It considers modern approaches to the solution of the spectral problem for real-value matrices with moderate dimensionality, such approaches being based on direct and inverse iterations (including iterations with displacement), as well as on the orthogonal transformation of the matrix to a diagonal or triangular form. The chapter describes the ideas on which these methods are based, provides derivation of formulas for the use in calculations, and describes specific algorithms which, in specified situations, allow one to solve partial and full eigenvalue problems.